Systematic Analysis of Posterior HOXA/HOXD Function in Mesenchymal Cells

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor of Philosophy
(Ph.D.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von

Dipl. Ing. Ivana Jerković

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr. Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen

1. Prof. Dr. Stefan Mundlos
2. Prof. Dr. Leonie Ringrose
3. Prof. Dr. Petra Seemann

Tag der mündlichen Prüfung: October 24, 2017

Summary

A multicellular organism developes from a one-cell zygote. During embryonic development, highly coordinated gene regulation enables cells to differentiate into various tissues and cell types. These processes are largely mediated by transcription factors (TFs). TFs are DNA binding proteins that impact the expression of other genes. This is accomplished by monomer, homo-, heterodimer, or through tethered TF-DNA binding. Cohorts of TFs often act in unison to precisely modulate the transcription of specific target genes. Homeobox (*Hox*) genes are essential developmental TFs that pattern the body plan along the anterior-posterior, and the appendages along the proximal-distal axes. During embryonic development, they are expressed in a nested fashion and perform partially redundant functions. Furthermore, i*n vitro* analysis of HOX-DNA binding identified overlapping sequence preferences that can be altered in the presence of specific cofactors. However, functional and biochemical investigation of genome-wide HOX binding patterns and delineation of individual *Hox* functions has been futile due to high protein homology, lack of specific antibodies, and nested expression domains. Due to lack of evidence, low *in vitro* HOX-DNA specificity and highly specific *Hox* functions have often been at odds, creating a so-called **Hox paradox.**

This study aimed to overcome these issues and investigate Hox paradox at the DNA binding level. For this, nine chicken limb-specific posterior *HOXA* and *HOXD* genes were FLAG-tagged, and overexpressed in limb-derived mesenchymal stem cells. In this system, native HOX cellular environment is preserved, and a unique epitope is created facilitating investigation of HOX-DNA binding and characterization of transcriptional programs induced by individual HOX.

In this system, individual HOX overexpression induces highly redundant regulatory programs. The extent of transcriptional redundancy within paralogy groups differs, with PG9 and PG13 inducing quite distinct, and PG10 and PG11 remarkably redundant regulatory programs. Likewise, HOX-DNA binding exhibits redundancy between and within paralogy groups. In line with the transcriptional redundancy, binding within PG10 and PG11 is more redundant than within PG9 and PG13. Furthermore, HOX bind DNA both directly and indirectly. Direct binding motifs discovered in this study are considerably different from the known monomer-like HOX motifs described *in vitro*. The change is mainly attributed to TALE and other yet undetermined cofactors. In contrast, indirect binding is responsible for a remarkable abundance

of HOX-DNA binding sites. Detailed inspection of HOX binding profiles identified subgrouping into two groups, Group 1 and Group 2. This subgrouping is linked to the abundance of indirect binding and partially to the CTCF mediated indirect binding. Moreover, binding sites co-occupied by both HOX and CTCF almost always exhibit additional Cohesin binding, indicating that these triple bound sites could have a role in the establishment and/or maintenance of local genome micro-architecture and chromatin remodeling. Finally, CTCF is confirmed as a novel HOX cofactor using two independent assays, implying that at least part of the contacts between HOX and CTCF are mediated through protein-protein interactions, whether directly or in a complex.

Zusammenfassung

Ein mehrzelliges Lebewesen entwickelt sich zum erwachsenen Organismus aus einer einzelligen Zygote. Eine hoch koordinierte Genregulation während der embryonalen Entwicklung ermöglicht die Differenzierung der Zygote in verschiedene Gewebe- und Zelltypen. Der Ablauf der Entwicklungsprozesse wird weitgehend durch Transkriptionsfaktoren (TFs) vermittelt. TFs sind DNA-bindende Proteine, die die Expression anderer Gene regulieren. Deren Bindung kann als Monomer, Hetero-, Homodimer oder als TF-DNA Bindungskomplexe erfolgen. Oft bilden TFs Komplexe miteinander und modulieren als solche die präzise Transkription spezifischer Zielgene.
Homöobox-Gene (*Hox*) sind wichtige TFs, die während der Entwicklung den Körperplan entlang der anteroposterioren Achse und die Anhänge entlang der proximodistalen Achse bestimmen. Während der embryonalen Entwicklung werden die *Hox*-Gene in verschachtelter Weise exprimiert und sind zum Teil redundant in ihren Funktionen. *In vitro* HOX-DNA-Bindungsanalysen ergaben überlappende Sequenzpräferenzen, die sich in Gegenwart von spezifischen Kofaktoren verändern.

Die funktionelle und biochemische Untersuchung von genomweiten HOX-Bindungsmustern und die Bestimmung der Funktionen einzelner *Hox*-Gene werden durch die hohe Proteinhomologie, unzureichend spezifische Antikörper und verschachtelte Expressionsdomänen erschwert und blieben bisher erfolglos. Der Mangel an Beweisen, die niedrigen HOX-DNA-Bindungsspezifität *in vitro* und die hochspezifischen Hox-Funktionen, die oft im Widerspruch stehen, führten zur Entstehung des so genannten Hox-Paradoxons.

Das Ziel dieser Arbeit bestand darin, diese Probleme zu überwinden und das Hox-Paradoxon in Bezug auf die DNA-Bindung zu untersuchen. Zu diesem Zweck habe ich neun Gliedmaßen-spezifische posteriore HOXA- und HOXD-Gene mit dem FLAG-Epitop markiert und diese in Gliedmaßen-mesenchymalen Stammzellkulturen überexprimiert. In diesem System wird die native HOX-Zellumgebung bewahrt. Dadurch wird die Untersuchung der HOX-DNA-Bindung und die Charakterisierung von Transkriptionsprogrammen, die durch einzelne HOX-TFs induziert werden, ermöglicht.

Die Überexpression der einzelnen HOX-TFs induziert hochgradig redundante Regulierungsprogramme. Das Ausmaß der Transkriptionsredundanz innerhalb der Paralogie-Gruppen (PG) unterscheidet sich: PG9 und PG13 induzieren deutlich unterschiedliche

Regulationsprogramme, während PG10 und PG11 stark redundant wirken. Die HOX-DNA-Bindung weist zugleich Redundanz zwischen und innerhalb von Paralogie-Gruppen auf. Im Einklang mit der Transkriptionsredundanz ist die HOX-DNA-Bindung zwischen PG10 und PG11 redundanter als zwischen PG9 und PG13. Genomweit können HOX sowohl direkt als auch indirekt an die DNA binden. Die in dieser Studie entdeckten direkten Bindungsmotive unterscheiden sich erheblich von den *in vitro* Monomer-ähnlichen HOX-Motiven. Diese Unterschiede sind hauptsächlich auf TALE und andere noch unbestimmte Kofaktoren zurückzuführen. Im Gegensatz dazu kommt es an einer bemerkenswert hohen Anzahl von HOX-DNA-Bindungsstellen zu einer indirekten Bindung. Im Besonderen das genomweite HOX-Bindungsprofil lässt sich unerwartet in zwei Gruppen unterteilen, Gruppe 1 und Gruppe 2. Diese Unterteilung steht mit der Menge der indirekten Bindungen und teilweise mit der CTCF-vermittelten indirekten Bindung in Zusammenhang.

Die Bindungsstellen, die sowohl von HOX als auch von CTCF erkannt werden, weisen sehr oft eine zusätzliche Cohesinbindung auf. Dies deutet darauf hin, dass diese dreifach gebundenen Stellen eine Rolle bei der Etablierung und/oder bei der Aufrechterhaltung der lokalen Genom-Mikroarchitektur und der Chromatin-Remodellierung spielen könnten.

Letztendlich wurde mittels zwei unabhängiger Experimente nachgewiesen, dass CTCF als neuartiger HOX-Kofaktor fungiert. Darüber hinaus besteht die Möglichkeit, dass zumindest ein Teil der Interaktionen zwischen HOX und CTCF indirekt in einem Komplex oder durch direkte Protein-Protein-Wechselwirkungen zustande kommt.

1 Introduction

Embryonic development is a complex and highly coordinated process where the embryo develops from the one-cell stage zygote to fully functional embryo with hundreds of specialized cell types. Therefore, all cells in the organism carry the same genetic information but still greatly differ from one another. This is accomplished through precise and timely gene expression which is often modulated by sequences and factors outside of the gene body itself. Finally, such accurate and titrated gene expression is monitored at multiple checkpoints at different levels; from transcription until translation.

1.1 Transcriptional Regulation

Transcriptional regulation is the first and most robust step in the control of gene expression. Simply put, transcriptional regulation is the regulation of timing and positioning of RNA polymerase II (RNA Pol II). This is achieved through regulation both in *cis* and *trans*, before and during the RNA transcription. While, generally, all genes are regulated on the transcriptional level, a specific group of genes, developmental genes, have to be expressed very specifically in space and time and therefore require extremely accurate transcriptional regulation. Below, I will focus on the most important aspects of transcriptional regulation of developmental gene expression.

1.1.1 Promoters

Promoters are gene proximal elements located immediately upstream of a transcription start site (TSS) that drive transcription. Promoters contain sequences that are essential for the recruitment of RNA Pol II and the formation of the preinitiation complex (PIC). The PIC is formed when general transcription factors (GTF) bind sequences upstream of the promoter triggering the nucleosome displacement and making the DNA accessible. Concomitantly, the chromatin state changes e.g. the histones get acetylated or methylated, thereby facilitating transcription (**Figure 1.1**) (Lee & Young 2000; Spitz & Furlong 2012; Kadonaga 2012). Formation of the PIC and the start of the transcription is guided by the elements present in the core promoter of every gene. Promoters of developmental genes contain a TATA box, the initiator, and the downstream

promoter element (DPE) that guide correct positioning of the RNA Pol II, PIC assembly, and instruct precise and accurate start site of the transcription (Roeder, 1996).

1.1.2 Enhancer Elements

Enhancers are *cis*-regulatory elements (CRE) that can activate a promoter and drive the transcription of a target gene. In contrast to the promoters, enhancers can act from both sides of the target TSS, in any orientation, and over extremely long linear distances (Banerji, Rusconi and Schaffner, 1981; Moreau *et al.*, 1981; Lee and Young, 2000; Lettice *et al.*, 2003; Sagai *et al.*, 2004). The discovery and localization of enhancers in animal genomes indicated that it is common for enhancers to be located far away from the cognate promoter, in the intergenic space or introns of other genes raising the question: How do the enhancers activate their target promoter and how do they achieve tissue specificity?

Answering this question is all but a trivial matter as tissue specific gene expression can be driven by one or more different enhancers. Furthermore, these enhancers can act in different tissues (tissue specific enhancers) or several enhancers can act together in the same tissue. Exemplarily, the *Shh* gene encodes a morphogen that is active in several tissues, including forebrain and limb

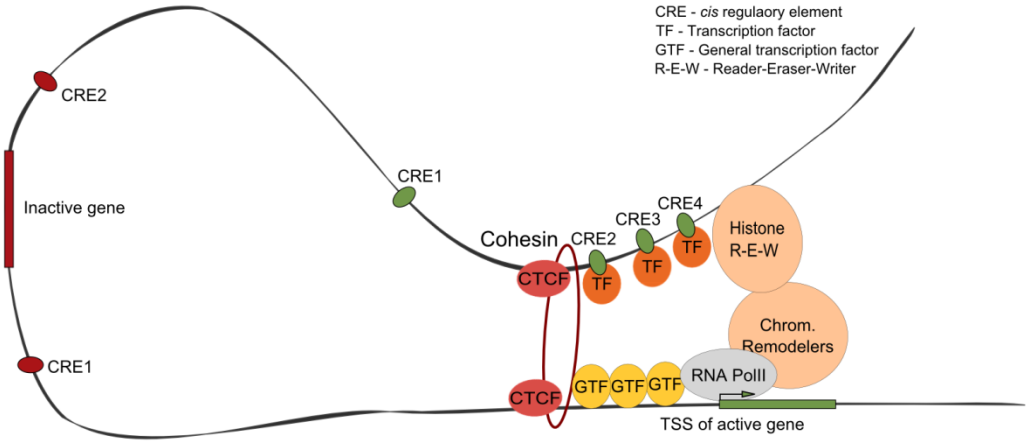


Figure 1.1 Schematic representation of the transcriptional initiation at an active gene.
Red rectangle and ovals represent inactive gene and *cis*-regulatory elements (CREs), respectively. Green rectangle and ovals represent active gene and *cis*-regulatory elements, respectively. In this case green CRE1 is a cis-regulatory element that is not active in the same tissue like CRE2, CRE3, and CRE4. In the scheme, active gene is contacted by the distal CREs that are bound by specific transcription factors (TFs). In order to come in the proximity to the promoter of the gene, the distal elements have to be physically moved closer, a process here referred to as looping. It is not fully clear how looping occurs at all positions in the genome but most often CTCF and Cohesin facilitate such conformational change, as depicted above. To allow for the transcription to occur at the promoter site general transcription factors (GTFs) have to be present as well as the histone readers, erasers and writers which will facilitate chromatin remodelers and facilitate start of the transcription via the RNA Pol II.

bud. In these two tissues, the expression of *Shh* is driven by tissue specific enhancers. In the limb, only by one enhancer, the ZRS, and in the forebrain by several different enhancers (Lettice *et al.*, 2003; Jeong *et al.*, 2006). Enhancers' tissue specificity is usually mediated by the presence or absence of the transcription factor binding sites (TFBS) at the enhancer body. Comparative analyses of enhancers between species indicated that some enhancers contain conserved transcription factor binding sites (TFBS) (Nobrega and Pennacchio, 2004; Kvon *et al.*, 2016). These TFBSs are instrumental for enhancers function as they allow them to operate in a modular fashion (**Figure 1.1**). Specifically, multiple TFBSs at the enhancer may provide malleability when activating the target gene in different tissue or time (Kulkarni and Arnosti, 2003; Arnosti and Kulkarni, 2005). Furthermore, a recent study has demonstrated that the TFBSs at enhancers are often sub-optimal. When these sub-optimal TFBSs were exchanged with high-affinity binding sites, target gene was both over- and misexpressed (Farley *et al.*, 2015). Therefore, TFBSs at enhancers are essential for the appropriate and precise gene expression.

1.1.3 Transcription Factors (TFs)

While promoters and enhancers are DNA elements that control gene expression in *cis*, transcription factors are DNA binding proteins that provide an additional layer of gene regulation in *trans*. Transcription factors can be general (GTF), specific[1], and structural. GTFs bind promoters whereas specific TFs bind enhancers and together they help bridge promoter and enhancer elements, form the PIC, and facilitate transcription (**Figure 1.1**). Additionally, structural TFs (e.g. CTCF) can arrange the chromatin geometry and in such manner impact the RNA pol II recruitment.

However, only specific TF binding at enhancer elements can activate enhancers and through this activation to modulate their target gene's activity (Deng *et al.*, 2012). This process is dependent on several factors, namely, the DNA sequence, protein-protein interactions, direct and indirect cooperativity, and TF concentration in the nucleus. So far there is no unifying model to describe the mechanism of enhancer activation as findings from different loci point towards different mechanisms. Currently, three different models are used to describe TF-mediated enhancers activation (Spitz and Furlong, 2012).

Research done on the interferon β enhancer suggests the logic of so-called **enhanceosome** model (Thanos and Maniatis, 1995; Merika and Thanos, 2001). Here, the enhancer contains

[1] From here on specific transcription factors are just referred to as transcription factors (TFs), unless specified otherwise.

transcription factor binding sites (TFBSs) that are precisely spaced following a strict grammar. According to this model, as TFs act cooperatively by binding to each other and using DNA as a scaffold if the binding of one TF is abrogated the enhancer is unable to produce a transcriptional output. Such enhancers can produce on/off difference in gene expression from one cell to another allowing for sharp gradients to form (Thanos and Maniatis, 1995; Merika and Thanos, 2001).

In contrast, the **billboard** model is characterized by more flexible grammar where not all motifs at the enhancer need to be bound for a transcriptional output to be produced. This more relaxed grammar allows for the same enhancer to be activated in different cellular environments and tissues while producing a similar output (Kulkarni and Arnosti, 2003; Arnosti and Kulkarni, 2005).

Finally, the investigation of the cardiac development uncovered a set of five TFs that co-bind a large number of enhancers and together control the expression of their target genes. Surprisingly, the sites occupied by these five TFs do not possess almost any grammar; underlying motifs are flexible and TF binding is highly dependent on protein-protein interactions. This model is known as "**collective transcription factor enhancer model**" (Junion *et al.*, 2012).

Three models described above mainly tackle direct cooperativity and motif grammar. However, in addition to the TFs themselves, there are other proteins (cofactors) that affect TF binding affility and dynamics. **Cofactors** can be specific (e.g. TALE-HOX) and general; either general coactivators (e.g. p300) or general corepressors (e.g. HDAC). General cofactors are cofactors that do not act on all TFs and often affect TF resting time, DNA affinity, or DNA accessibility (Merika *et al.*, 1998; Mann, Lelli and Joshi, 2009). Conversely, specific cofactors are targeting only a specific suggroup of TFs and can be tethering or sequence specific. One of the best known examples of specific tethering cofactors comes from the β-globin locus where distal locus control region (LCR) contacts β-globin promoter activating transcription. Importnantly, both, β-globin promoter and LCR are directly bound by the GATA1 transcription factor on top of which a an associated molecule, Ldb1, binds causing Ldb1 self-dimerization, looping LCR to β-globin promoter, and activates transcription (Deng *et al.*, 2012). On the other hand, sequence specific cofactors heterodimerize with TFs and together co-bind to the DNA. Best known example of sequence specific cofactors is the HOX-TALE association. More specifically, Scr is a *Drosophila* Hox transcription factor which requires an association with Exd in order to bind a subset of its binding sites. Only in the case of the homodimerization, the Scr-Exd complex recognizes the

low-affinity conglomerate sites that Scr alone would not be able to recognize, therby facilitating the regulation of essential target genes (Crocker *et al.*, 2015). Therefore, in order to produce precise and timely gene regulation the presence and coordination of GTFs, structural TFs, and both, tethering and sequence specific cofactors is required and necessary.

1.1.4 Genome Architecture

In addition to specific TFs and GTFs, structural TFs impact gene expression by altering the chromatin geometry. Structural TFs are a class of DNA-binding proteins that can assure the accuracy of gene expression by affecting the genome folding and some affect partitioning the genome into self-associating isolated units, topologically associated domains – TADs, characterized by more frequent DNA-DNA contacts within than outside of the domain (**Figure 1.1** and **Figure 1.2**). Recent studies uncovered that target genes and their enhancers mainly occupy chromatin belonging to the same TAD, indicating that genome architecture has a functional component to it. These units are usually several hundred kilobases up to a megabase in size and are flanked by a DNA stretch with insulating power separating two adjacent TADs (**Figure 1.2**). These DNA stretches are termed boundaries. Boundaries encompass DNA of tens of kilobases in size and are often associated with structural TFs like CCTC-binding factor (CTCF), Cohesin, and/or strong transcription (Dixon *et al.*, 2012; Nora *et al.*, 2012, 2016; Phillips-Cremins and Corces, 2013; Zuin *et al.*, 2013; Schwarzer *et al.*, 2016). In addition to preferential boundary localization, CTCF together with Cohesin can mediate "looping" between two pieces of DNA within the TAD (e.g. enhancer-promoter contacts) (Zuin *et al.*, 2013; Rao *et al.*, 2014). Furthermore, it has been shown that the CTCF binding motif orientation plays a role in the formation of DNA loops. Specifically, when mediating loop formation two CTCF binding sites are almost exclusively in convergent orientation (de Wit *et al.*, 2015; Guo *et al.*, 2015).

Additional to structural TFs, a recent study demonstrated that specific TFs can affect local TAD architecture as well. Beccari et al. (2016) demonstrated HOX13 proteins are essential for the switch from early to late limb regulation at the HoxD locus. More specifically, HoxD cluster is located directly at the boundary of two TADs and on the either side of the HoxD cluster are HOX enhancers that belong to different TADs, early enhancers to the tTAD and late enhancers to cTAD. During early limb patterning only early enhancers (tTAD) are active whereas during late patterning only late enhancers are active (cTAD). In *Hox13* loss-of-function mouse the switch from early regulation (cTAD) to late regulation (tTAD) fails and mutant animals do not properly form the autopod. However, although these experiments clearly demonstrate that

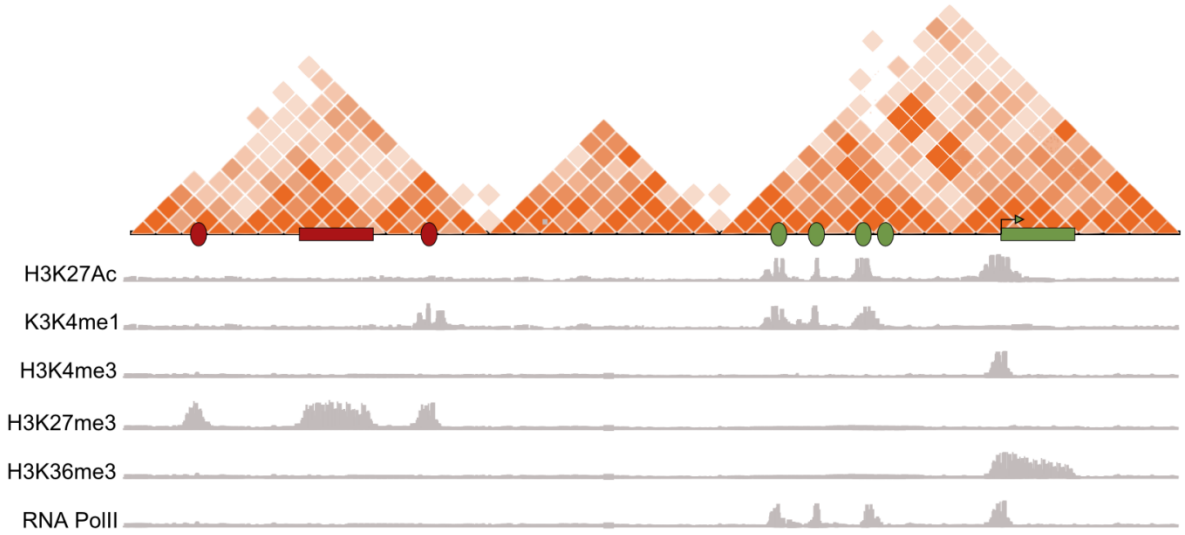


Figure 1.2 Schematic representation of genomic architecture at an active and an inactive gene and their accompanying histone modifications.
Red matrix on top represents a Hi-C map over the region of the genome. Red triangles are regions that predominantly interact more frequent with each other, so called, topologically associating domains (TADs). Under TADs there are either active or inactive genes where their respective cis-regulatory elements (CRE) either do or don't contact the genes. Histone modifications mark genes and cis-regulatory elements. They bear marks as follows: 1) Active gene-H3K4me3, H3K27Ac, HK36me3 (over the gene body) and RNA PolII, 2) Inactive gene-H3K27me3, 3) Active CRE-H3K27Ac, H3K4me2, RNA PolII, or none of the mentioned if the CRE is not active in the present tissue (green CRE closest to the active gene), and 4) Inactive CRE-H3K27me3 and if enhancer is poised additionally, H3K4me1. Adapted from (Ong and Corces, 2014).

HOX13 proteins govern the switch between early and late limb regulation, the exact mechanism of this TAD switch is still unclear.

Taken together, recent discoveries clearly demonstrated that TFs utilize chromatin structure in order to modulate gene expression. In such manner, structural TFs (CTCF) mold the global genome architecture to restrict transcription to functional units, and specific TFs further manipulate local micro-architecture to instigate precise transcriptional output.

1.2 *Hox* Genes

Homeobox (*Hox*) genes are a gene family coding for transcription factors (TFs). *Hox* genes are highly conserved in all bilaterians where they direct patterning of the body plan along the anterior-posterior (A-P) and patterning of appendages along the proximal-distal (P-D) axis (Lewis, 1978; Akam, 1989). They were first identified during a large mutation screen in *Drosophila*, as they caused very peculiar phenotypes. Specifically, loss-of-function and gain-of-function

mutations in *Antp* and *Ubx* genes cause the transformation of one body segment into the likeness of another, antennae-to-leg and haltere-to-wing, respectively (Lewis, 1978), a phenomenon termed Homeotic transformation.

All *Hox* genes are characterized by three properties. First, by the **clustered nature of the *Hox* genomic organization** (Pascual-Anaya *et al.*, 2013). Second, *Hox* genes are **expressed in a nested and overlapping pattern** according to their relative genomic position on the chromosome, both in time (e.g. *Hox1* is expressed first) and space (e.g. *Hox1* is expressed most anterior). This pattern of expression is referred to as spatiotemporal collinearity. Third, they contain a conserved DNA binding domain, **the Homeodomain (HD)**. This is a 60 amino acid (aa) sequence (with the exception of TALE (Three Amino Acid Loop Extension) proteins which have 63aa Homeodomain) located at the C-terminal part of all Homeoproteins (Gehring *et al.*, 1994). Below, I will focus on vertebrate *Hox* genes during limb development and on the properties of HOX-DNA binding.

1.2.1 Vertebrate *Hox* Genes

In tetrapod vertebrates there are 39 *Hox* genes organized in four clusters on four different chromosomes. In higher animals, *Hox* genes increased in number with the series of tandem and whole genome duplications from a single — proto-Hox — cluster. After cluster duplications, the evolutionary pressure on the newly duplicated genes temporarily reduced creating "window of evolvability" (Wagner, Amemiya and Ruddle, 2003). Due to increased susceptibility to genomic changes *Hox* genes were more likely to adopt a new function or to disappear. This process is visible in higher organisms where in spite of the fact that four Hox clusters are present (HoxA, HoxB, HoxC, and HoxD) only three paralogy groups (PG) contain all four paralogues, those are PG4, PG9, and PG13 (**Figure 1.3)** (Wagner, Amemiya and Ruddle, 2003)**.** Due to these processes, today, there is more pronounced functional divergence within the paralogy group that contains all four paralogues than within paralogy groups missing one or more paralogues. This will be discussed in more detail in **Chapter 4.4** and **Chapter 5.1.2**.

1.2.2 *Hox* Function in Developing Limb Bud

Vertebrate *Hox* genes are homologous and as such, functionally partially redundant. Just like in flys, they are essential for patterning of the body plan along the A-P axis, but also for development of other novel structures like: metanephric kidneys, lungs, intestines, genitalia, and limb buds (Akam 1989; Gong et al. 2007; Di-Poi et al. 2007; Zakany & Duboule 2007; Kondo et al. 1997; Featherstone et al. 1988). In the developing limb bud, almost exclusively, only nine posterior *Hox*a and *Hox*d genes are expressed (*Hoxa9-13* and *Hoxd9-13*) in overlapping patterns which allow them to build the limb structures in a combinatorial fashion (**Figure 1.4A**, **B**, and **C**). In such way, Hox9 and Hox10 paralogues mainly pattern zeugopod, Hox11 stylopod, and Hoxd12 and Hox13 paralogues pattern autopod (Small & Potter 1993; Davis & Capecchi 1994; Fromental-Ramain et al. 1996; Kmita et al. 2002; Kmita et al. 2005; Duboule 2007).

So far, geneticists used mainly loss-of-function and knock-out experiments to decipher individual *Hox* gene functions. Knock-out mice of, *Hoxa11* or *Hoxd11* gene exhibit only mildly misshapen

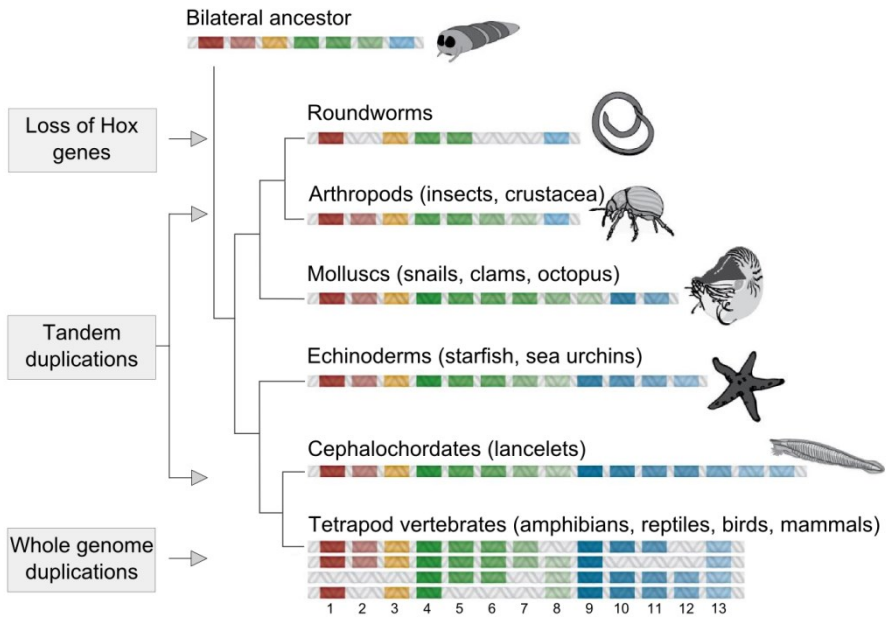


Figure 1.3 Phylogenetic tree showing emergence from the proto-Hox cluster to four Hox clusters in vertebrates.
Hox have emerged from a proto-*Hox* cluster in bilateral ancestor that was subjected to changes throughout evolution. In roundworms some genes get lost. In arthropods, molluscs, echinoderms and cephalochordates a series of different tandem duplications multiply the genes. Finally, in tetrapods with the two rounds of whole genome duplications four Hox clusters emerge with total of 39 genes. Adapted from (Genetic Science Learning Center, 2016).

ulna and radius. However, a double *Hoxa11/d11* knock-out mutant mouse presented a striking reduction in the size of ulna and radius that was not present in the knock-out of either *Hoxa11* or *Hoxg11* alone (Davis *et al.*, 1995). These experiments demonstrated both the importance of the Hox11 paralogues for the development of zeugopod and their individual functional redundancies. Similarly, *Hoxa10/d10* loss-of-function mice exhibited a noticeable reduction in femur indicating the importance of the Hox10 paralogues for the stylopod patterning (Wellik & Capecchi 2003; Fromental-Ramain et al. 1996) whereas the *Hoxa13/d13* loss-of-function mice exhibit a completely abrogated autopod development revealing *Hox13* genes as master regulators of autopod development (Dolle et al. 1993; Fromental-Ramain et al. 1996). Additionally, these findings were substantiated with an elegant series of genetic experiments where and entire HoxA and/or HoxD clusters were deleted (**Figure 1.5**). Surprisingly, the deletion of a single cluster, either HoxA or HoxD, lead only to very mild phenotypic consequences for the developing limb. However, upon the deletion of both clusters (HoxA and HoxD) limb bud development halts and only the most proximal structures of the limb develop (scapula) revealing the extent of redundancy between these genes (Kmita *et al.*, 2002; Marie Kmita *et al.*, 2005).

Importantly, Hox functional redundancy extends beyond their own paralogy group as some neighboring genes are found to be partially redundant as well (Zakany & Duboule 2007; Woltering & Duboule 2010). Exemplarily, the experiments where the Homeodomain of one *Hox* gene was swapped for the Homeodomain of another *Hox* gene have demonstrated that some Homeodomains are interchangeable to no phenotypic expense whereas others are not (Zhao and Potter, 2002). This was further elaborated and extended with a series of *in vivo* experiments. In those studies, a deletion of *Hoxd13* gene caused *Hoxd12* gene to take the terminal position in the HoxD cluster and *Hoxd12* gene to adopt a *Hoxd13* pattern of expression. Surprisingly, however, the phenotypic consequence of this perturbation was barely noticeable (Kmita *et al.*, 2002). Conversely, when *Hoxd13* was mutated and the HOXD13 protein rendered inactive, but not removed from the genome, *Hoxd12* expression was restricted to its own endogenous domain and the mice exhibited polydactyly which was not present in *Hoxd13* deletion. Next, the same logic was followed for the experiments where *Hoxd13* and *Hoxd12* were deleted causing the *Hoxd11* gene to relocate to the *Hoxd13* position. Expectedly, this caused *Hoxd11* to adopt *Hoxd13* gene expression domain. Surprisingly, however, this perturbation caused polydactyly which was even more severe than polydactyly present in the *Hoxd13* loss-of-function mice (Kmita *et al.*, 2002).

Together, these genetic experiments indicate two important characteristics of *Hox* gene function in the limb bud, nested expression pattern and redundancy. Nested, but distinct expression domains are essential and necessary for fine tuning of patterning and differentiation. However, even if expressed in the same expression domains, not all *Hox* genes are equivalently redundant, and the extent of redundancy, although most prominently present in the individual paralogy groups, extends to the neighboring genes as well (e.g. *Hoxa13-Hoxd13* but also *Hoxd12-Hoxd13*).

1.2.3 HOX-DNA Binding

HOX proteins' main function is DNA binding that is carried out by the most conserved part of the protein, the Homeodomain. The Homeodomain forms three alpha helices connected with turns where first two helices mainly stabilize the binding and a third, the most C-terminally positioned, is the recognition helix responsible for most of the sequence recognition on the DNA. HOX-DNA binding is mainly mediated through amino acids at the positions 47, 50, 51 and 54 of the Homeodomain which code for the Ile, Gln, Asn, and Met (Met can be replaced by Val in HOX13, in Aves), respectively (Gehring *et al.*, 1994). Furthermore, amino acids mediating

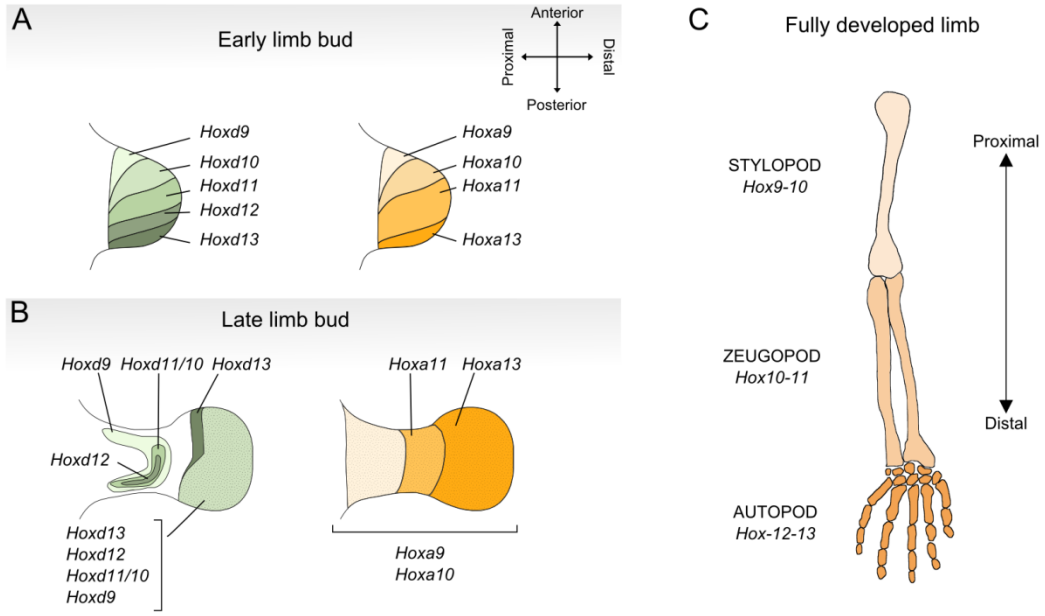


Figure 1.4 Schematic representation of *Hoxa* and *Hoxd* expression in early and late limb bud, and their contribution to the limb morphology.
A) *Hoxa* and *Hoxd* are expressed in limb-covering, nested domains in the early limb bud. B) In late limb bud this nested pattern continues with the exception that *Hoxd* genes are not expressed in the future wrist area. C) Fully developed limb arises with the anatomically specific contributions of *Hox* genes. PG9 and PG10 help shape the zeugopod, PG10 and PG11 the stylopod and PG13 and *Hoxd12,* almost exclusively, shape the autopod (right panel). According to (Zakany & Duboule 2007).

this HOX-DNA contact are extremely conserved in all bilateria emphasizing the preservation of the DNA binding mechanism in this protein family. However, despite high conservation of direct binding specificity driven by Homeodomain (due to the high protein homology), HOX proteins exhibit diverse function *in vivo* (Mann, Lelli and Joshi, 2009). This discrepancy between HOX biochemical redundancy and functional specificity has been termed **HOX Paradox.**

Several large-scale studies examined Homeodomain specificity *in vitro*, both in fruit fly and mouse (Noyes et al. 2008; Berger et al. 2008; Slattery et al. 2011; Jolma et al. 2013; Jolma et al. 2015). In mouse, posterior paralogues bind four classes of motifs: HOXA9/A10/D10 TFs bind the [C/T][A/C]ATAAA; HOXA11/D11 bind CTCGTAAA; and HOXA13/D13 bind, both CTCGTAAA and CCAATAAA motifs (Berger et al. 2008; Zhang et al. 2011; Turner et al. 2014). These findings indicate significant overlap between different HOX binding specificities while it is known from various genetic experiments that they perform distinct roles in the developing embryo. Therefore, full *in vivo* binding specificity must rely on additional cofactors, both co-binding and perhaps tethering the HOX proteins, in order to confer the appropriate HOX function(s). Very little is known about HOX binding *in vivo* and about their cofactors. Up-to-date most of what we know about HOX-cofactor binding comes from the best known HOX cofactors, Three Amino Acid Loop Extension (TALE) proteins. They are Homeodomain family of TFs conserved from plants to animals and in Vertebrates come in two subfamilies PBC (Pbx1-4) and MEIS/PREP (Meis1-3 and Prep1-2) (Mukherjee and Bürglin, 2007). HOX and TALE proteins co-bind and alter sequence specificity before binding to DNA (LaRonde-LeBlanc and Wolberger, 2003). Specifically, HOX contact PBC proteins through N-terminally positioned Hexapeptide (HX) motif that consists of YPWM amino acids and with HX contacts PBC's Homeodomain. TALE TFs have three additional amino acids in their Homeodomain, from where they get their name, which PBC proteins use to make a hydrophobic cavity where Hox Hexapeptide (HX) motif binds to PBC's Tryptophan (W) (Mann, Lelli and Joshi, 2009). Additionally to the HX motif, on the C-terminal part of the HOX Homeodomain, there are short motifs that drive specific HOX-PBC contacts, collectively known as SPIM sequences (specific PBC interaction motifs). They drive specific interactions only in a subset of HOX proteins as they are not particularly well conserved and are not present in all HOX proteins.

Taken together, aforementioned *in vitro* analyses focused mainly on direct Homeodomain-DNA binding specificity. However, because of the large-scale scope of these analyses they did not focus on specificities of cofactor driven HOX binding and therefore could not account for the actual *in vivo* specificity of these complexes. Indeed, the study using SELEX-seq identified significant

changes in sequence specificity occurring once HOX co-bound with TALE proteins (Slattery *et al.*, 2011). This phenomenon where the sequence specificity is changed upon the binding with another protein is known as **latent specificity** and at least partially explains the Hox Paradox. Furthermore, an *in vivo* study examining multiple HOX binding sites in *Drosophila* found that high-affinity binding sites do not entail biological importance of these sites, but rather the opposite. In other words, the more specific the binding site (e.g. only one HOX paralogue can bind, or only HOX paralogue with a cofactor), the lower affinity of the site (Crocker *et al.*, 2015). It is not entirely clear why the specificity inversely correlates with the affinity. However, it is clear that common (i.e. where any HOX-TF can bind) consensus sites are high-affinity sites that often are not of utmost biological relevance, in contrast to the hybrid HOX-cofactor sites.

In this project, the latent specificity was encountered as well, indicating that low-affinity, biologically important sites are common and a general feature of HOX binding. These findings will be discussed in detail in **Chapter 4.5.4**, **Chapter 5.2.1**, and **Chapter 5.2.3**.

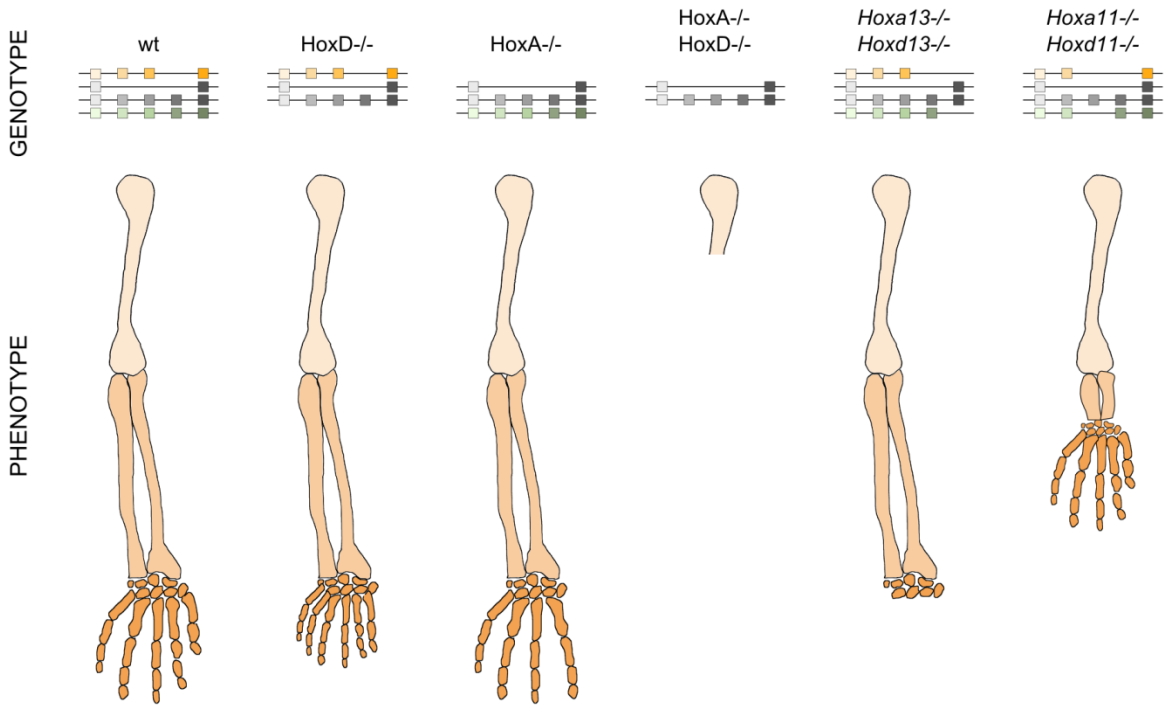


Figure 1.5 Schematic representation of *Hoxa* and *Hoxd* deletion experiments and their contribution to the limb related phenotype.
A wild-type and five deletion experiments with entire HoxD, entire HoxA, entire HoxD and HoxA, PG13, and PG11 deletions, respectively, with their related limb phenotypes depicted under the genotype. According to (Zakany & Duboule 2007).

Additional to binding specificity conferred by specific cofactors, binding can be influenced by general cofactors as well. Specifically, HOX were shown to associate with CBP/P300 proteins to more efficiently regulate its expression. CBP/p300 proteins contain a histone acetyltransferase (HAT) domain that can normally acetylate histones and help activate the target gene. Interestingly, however, with HOX, CBP acts as a corepressor rather than coactivator. CBP binds with HOX in a complex before it reaches the DNA which prevents HOX-DNA binding. Additionally, when HOX-CBP form a complex CBP's HAT activity is blocked leading to downregulation and not upregulation of target genes (Shen *et al.*, 2001). Finally, one of the very few evidence of tethered HOX binding was demonstrated for the HOXD13 mutated protein, where HOXD13 DNA binding activity has been abolished. Interestingly, such HOXD13 mutant was able to upregulate some targets much like its wild-type counterpart. However, this construct was unable to negatively regulate some of its own targets (Williams, Williams, Kuick, *et al.*, 2005). Together, these two examples demonstrate that subsets of binding sites in the genome undoubtedly rely on the mechanisms that circumvent both the direct and TALE modified direct binding. However it is still unclear how abundant are indirect and tethered binding and how important they are for the HOX function.

In conclusion, HOX binding sites are extremely heterogeneous. Source of such heterogeneity are various binding modules of HOX-TFs as they utilize both specific and general cofactors. The combination of nested expression patterns, specific cell type cofactors, and specific HOX cofactors allow transformation of initially unspecific Homeodomain recognition motif into a precise output. The knowledge and investigation of this complex regulatory input will permit a better understanding of Hox paradox and help elucidate how HOX-TFs exert their function.

2 Aim of the Thesis

Homeobox (*Hox*) genes are essential developmental transcription factors that pattern the animal body. They are expressed in an overlapping pattern along the trunk, in the developing appendages, and organs. In the developing limb bud, mainly posterior nine *Hoxa* and *Hoxd* genes are expressed. Posterior HOXA and HOXD proteins are highly homologous and function in a partially redundant manner. *In vitro* examination of the HOX-DNA binding preferences uncovered remarkably similar sequence specificity. The inconsistency between nearly identical *in vitro* binding and specific Hox TF function is at the core of problem known as a Hox paradox.

To address this paradox, it is necessary to study HOX-DNA binding *in vivo* and in a Hox native environment. However, overlapping Hox genes expression patterns, high protein homology, and lack of highly specific antibodies prevented these analyses for years. The aim of this study is to address the inconsistencies known as a Hox paradox by exploring HOX-DNA binding *in vivo*.

To do so, this study utilized a primary cell-culture system to provide a native Hox cellular environment. Additionally, a unique epitope was fused with the *Hox* genes allowing discrimination between individual HOX proteins. Using this system together with a combination of genomics, genetics, and biochemical approaches the genome-wide binding profiles and individually induced transcriptional programs of nine posterior HOXA and HOXD TFs were investigated. Finally, the direct and indirect HOX-DNA binding were uncovered as HOX binding modes, where the focus was on the indirect, noncanonical HOX binding and on cofactor(s) that aid this process.

3 Material and Methods

3.1 Materials

3.1.1 Chemicals

Unless stated otherwise, chemicals were obtained from Merck (Darmstadt), Roth (Karlsruhe) or Sigma-Aldrich (Hamburg, Seelze, Schnelldorf, and Steinheim) in analytical grade quality.

3.1.2 Buffers

Common buffers and solutions were prepared according to Sambrook & Russell (2001).

Buffers for Chromatin Immunoprecipitation:

Lysis Buffer 1:	50mM HEPES-KOH, pH7.5; 140mM NaCl; 1mM EDTA; 10% Glycerol; 0,5% NP-40; 0,25% Triton X-100; Protease Inhibitors (Roche complete, add fresh)
Lysis Buffer 2:	10mM Tris-HCl, pH 8.0; 200mM NaCl; 1mM EDTA; 0.5mM EGTA; Protease Inhibitors (Roche complete, add fresh)
Lysis Buffer 3:	10mM Tris-HCl, pH 8.0; 100mM NaCl; 1mM EDTA; 0.5mM EGTA; 0,1% Na-Deoxycholate; 0,5% N-Laurylsarcosine; Protease Inhibitors (Roche complete, add fresh)
RIPA (Wash Buffer) :	50mM HEPES-KOH, pKa 7.55; 500mM LiCl; 1mM EDTA; 1,0% NP-40; 0,7% Na-Deoxycholate; Protease Inhibitors (Roche complete, add fresh)
TE-NaCl:	10mM Tris-HCl, pH 8.0; 1mM EDTA; 50mM NaCl, Protease Inhibitors (Roche complete, add fresh)
ChIP-Elution Buffer:	50 mM Tris-HCl, pH8.0; 10mM EDTA, 1.0% SDS

Buffers for Co-Immunoprecipitation:

Low-salt Buffer:	10 mM HEPES-KOH, pH 7.9; 1.5 mM MgCl2; 10 mM KCl; 0.5 mM DTT; 0.2 mM PMSF; Protease Inhibitors (Roche complete, add fresh)
High-salt Buffer:	20 mM HEPES-KOH, pH 7.9; 25% Glycerin; 420 mM NaCl; 1.5 mM MgCl2; 0.5 mM DTT; Protease Inhibitors (Roche complete, add fresh)
Blocking Buffer:	0.25% BSA in Wash Buffer 105
BS3 Conjugation Buffer:	150mM NaCl; 20mM NaH2PO4
BS3 Stock Solution:	2 mg of BS3 (Thermo Fisher Scientific, #21585) in 35 µL BS3 Conjugation Buffer
BS3 Working Solution:	5% BS3 Stock Solution in BS3 Conjugation Buffer
Quenching Buffer:	Tris-HCl, pH 7.5
IP Buffer:	20mM HEPES-KOH, pH 7.9; 1.5mM MgCl2; 0.5mM DTT; 105mM NaCl; 6.25% Glycerin; sterile filter; Protease Inhibitors (Roche complete, add fresh)
Wash Buffer 105:	20mM HEPES-KOH pH 7.9, 1.5mM MgCl2, 0.5mM DTT, 105mM NaCl, sterile filter; Protease Inhibitors (Roche complete, add fresh)
Wash Buffer 150:	20mM HEPES-KOH pH 7.9, 1.5mM MgCl2, 0.5mM DTT, 150mM NaCl, sterile filter; Protease Inhibitors (Roche complete, add fresh)
Co-IP Elution Buffer:	50mM Tris; 1% SDS; 10mM EDTA

Buffers for Proximitiy Ligation Assay:

PLA Blocking Solution (TSA):	10% horse serum; 0,5% PerkinElmer blocking reagent (#FP1020); 0.01% Triton-X-100 in 1x DPBS
PLA Antibody Diluent:	10% horse serum in 1x fresh DPBST
PLA Buffer A:	0.01 M Tris, 0.15 M; NaCl; 0.05% Tween 20; sterile filter
PLA Buffer B:	0.2 M Tris; 0.1 M NaCl; sterile filter

3.1.3 Media

Table 3.1 Cell Culture Media

Name of the media	Supplier and catalog number
DMEM; 4.5g/L Glucose	Lonza #12-614F
DMEM; 1g/L Glucose	Lonza #12-707F
DMEM Ham‘s F-12 (1:1)	Biochrom #F4815

3.1.4 Antibodies

Table 3.2 Antibodies

Antibody name	Supplier and catalog number
Mouse αFLAGM2	Sigma-Aldrich, #F1804-5MG
Mouse αHA (HA11.1 epitope tag)	BioLegend, #901501
Rabbit αCTCF	Active Motif, # 61311
Rabbit αRAD21	Abcam, #ab992
Rabbit αH3K4me3	Millipore, #07-473
Rabbit αH3K27me3	Millipore, #07-449
Goat αmouse IgG HRP	Millipore, #12-349
Normal Rabbit IgG	Cell Signaling Technology, #2729

Protein G magnetic beads were purchased from Invitrogen (Dynabeads, #100.04D).

3.1.5 Enzymes

Restriction enzymes were purchased from NEB (Frankfurt) or MBI-Fermentas (St. Leon-Roth). Taq- and Pfu-DNA-polymerases were produced in-house (A.C. Stiege). Phusion DNA-Polymerase was purchased from NEB, T4-ligase, and Polymerase from MBI- Fermentas, and RNase A (# R4875) and Proteinase K (# P2308) from Sigma-Aldrich.

3.1.6 Primers

Table 3.3 *HOX* and *CTCF* Amplification and Cloning Primers

Name	Sequence (5‘→3‘)
chHoxA9_F	CGCGTCTCCCATGTCGGCCCCCGGGACCCTC
chHoxA9_R	CGACTAGTTCATTCGTCCTTCGCTCGGTCTTTGTTGATTTTCTTC
chHoxA10_F	CCCGTCTCCCATGTCATGCTCCGAGAGCCCGGC
chHoxA10_R	CGAACTAGTTCAAGAGAAATTAAAGTTGGCTGTGAGCTCCC
chHoxA11_F	CCTCATGATGGATTTTGATGAGCGTGTTCCTT
chHoxA11_R	CTAACTAGTTTAAAGTAGTGGATTAGCTGAGTAATATTGTAA
chHoxA13_F	GCACATGTTCCTCTACGACAACAGCCTGGATGAG
chHoxA13_R	TAACTAGTTAACTGGTCGTCTTCAATTTGTTGATGAC
chHoxD9_F	TCATGATGTCGTCTAGTGGCACCATAAG
chHoxD9_R	CCACTAGTTAGTCTCCTTTATTGCCT
chHoxD10_F	CCACATGTCCTTTCCCAACAGCTC
chHoxD10_R	CCACTAGTTTAGGAGAAGGTCAGATTAG
chHoxD11_F	CCTCATGACCGAGTTTGACGATTGCAGTCACG
chHoxD11_R	CCACTAGTCAAAACAAGGGATTTCCAGTGAAGTATTGG
chHoxD12_F	CCCGTCTCACATGTGTGATCGCAGTCTCTACAGATCTGGCTAC
chHoxD12_R	GCACTAGTACATAGAGAGCGCCTGCTCGCG
chHoxD13_F	CGGAAGACGACATGGACGGACTGCGCGGCG
chHoxD13_R	CGACTAGTCAAGAAACGTTGTCTTTCAGTTTGGAGAC
chCTCF_gibFL_F	AGGATGACGATGACAAGTCCATGGAAGGTGAAGCAGTTGAAGCCA
chCTCF_gibFL_R	TGGCTTCAACTGCTTCACCTTCCATGGACTTGTCATCGTCATCCT
chCTCF_gibHA_F	CCGATTACGCCAGCAAGTCCATGGAAGGTGAAGCAGTTGAAGCCA
chCTCF_gibHA_R	TGGCTTCAACTGCTTCACCTTCCATGGACTTGCTGGCGTAATCGG

Table 3.4 Primers for Cloning of *HOXA10* Deletion Constructs

Name	Sequence (5‘→3‘)
F_chA10_Δ1-60	ATGCTGTTCCCCGTCCTGGGCAA
R_chA10_Δ1-60	GGACTTGTCATCGTCATCCT
F_chA10_Δ60-120	TCCTACTGCCTCTATGACTC
R_chA10_Δ60-120	TCCGCAGCCCTGCAGCCCGT
F_chA10_Δ120-180	GCCGGGACGGCCCCCTTCGC
R_chA10_Δ120-180	GCTCTCCTCCTTAATGTTTT
F_chA10_Δ180-240	CCCGCGCCGTCGGAGGGCAG
R_chA10_Δ180-240	CGGCCCCGCGCCGTAGCCCT
F_chA10_Δ240-270	AAAAGTGGCCGAAAGAAACG
R_chA10_Δ240-270	GGAGAGCTCCTCGGCGGTGG
F_chA10_Δ170-200	CCGCCCGTGCCGGAGGCGGG
R_chA10_Δ170-200	GTAAGCCTGGGAGAGGCGGA
F_chA10_Δ270-347	TGAACTAGTTACGATGCGATGTA
R_chA10_Δ270-347	TGAACTAGTTACGATGCGATGTA

Table 3.5 Sequencing Primers

Name	Sequence (5‘→3‘)
chHoxA9_seq_F	ACCTACCAGCAGGCATTACG
chHoxA9_seq_R	CGGTCCCTGGTGAGATACAT
chHoxA10_seq_F1	AGGTACTTCGCAAAGCATGG
chHoxA10_seq_R1	CTGGGAGAGGCGGAAGTAG
chHoxA10_seq_F2	AGCCCGTAGGCAATTCAAA
chHoxA10_seq_R2	AGTTTCATTCTGCGGTTCTGA
chHoxA11_seq_F	CCAACGTCTCCTCCAATTTCT
chHoxA11_seq_R	TGATATTTGGTGTAGGGGCATCT
chHoxA13_seq_F	TACATGGACACGTCGGTCTC
chHoxA13_seq_R	ACCTTTGTGTAGGGCACTCG
chHoxD9_seq_F	CGCCACTACGGGATAAAGC
chHoxD10_seq_F	CAGACGTCCCTTCCTACCAG
chHoxD10_seq_R	CAGGCAGCTCCTCTCGTCTT
chHoxD11_seq_F	CGCCTCCAACTTCTACGG
chHoxD11_seq_R	TTGAAGAAAAACTCACGTTCCA
chHoxD12_seq_F	GAGGAAAGATGCAGGCAGAG
chHoxD12_seq_R	TTCTTTCCTCTTCTGTCGGTTA
chHoxD13_seq_F	CATGGACGTCTCCAGTCTGA
chHoxD13_seq_R	GCTTGGTGTAGGGCACTCTC

chHoxA13_seq_F2	GAACTCGAAAGGGAATACGC
chHoxA13_seq_R2	CATGTACTTGTCGGCGAAGG
chHoxA13_seq_R3	CTCGCTCGACGTGTAGGC
chHoxD11_seq_F2	GCAGTTGTCCAGAATGCTCA
chHoxA9_seq_F2ex	CAGGCTCCTCAACCTCACC
chHoxA10_seq_F2ex	GAGCGTCGCCTAGAGATCAG
chHoxA11_seq_F2ex	AACAAAGAGAAACGCCTCCA
chHoxA13_seq_F2ex	AGACAAACGGCGGAGGATA
chHoxD9_seq_F2ex	GAGCGAGGGAAAGAAAGGAA
chHoxD10_seq_F2ex	CGCGGCTAAAGTCTCTCAAG
chHoxD11_seq_F2ex	TGGAACGTGAGTTTTTCTTCAA
chHoxD12_seq_F2ex	CGAAAGAAACGGAAACCGTA
chHoxD13_seq_F2ex	GAGTGCCCTACACCAAGCTC
chHoxD13seq_F3	CAGTGCCGTAACTTTCTCTC
chHoxD13seq_F4	CTCTCCTCGCCCGTTTTC
chHoxD13seq_R2	TCCTGGGTACATAGACATGG
chHoxD13seq_R3	TTTGTATAGCCCTGGTAGGA
chHoxD13seq_R4	CAACGGATACTATAGCTGCA
ch_Hoxd13_R5	AGCCCGGGCAGTCCTTG
ch_CTCF_Fnorm.	ATGGAAGGTGAAGCAGTTGAAGC
ch_CTCF_Rnorm.	TCACCGGTCCATCATGCT
ch_CTCF_F1_seq	AGGCTACGGTGGATGATACG
ch_CTCF_F2_seq	CATTCCAGTGTGAACTGTGC
ch_CTCF_F3_seq	GCAGAAGCACACGGAGAAC
ch_CTCF_F4_seq	GAGACAAAGAAGGGCAAACG
ch_CTCF_R1_seq	GCTATATGGCAAGCCTCGTC
ch_CTCF_R2_seq	TTTGGCTGGTGGCTGATAGT
ch_CTCF_R3_seq	TGTGGCGTTTCAATTTGCTA
RCAS_FLAG_F	GCCGACCACCATGTCTGAC
RCAS 5´seq	TCCATCAGCTACCACACGGAA

3.1.7 Kits

All below listed kits were used according to manufacturer's recommendation.

Table 3.6 Molecular Biology Kits

Kit name	Supplier
NucleoSpin Plasmid	Macherey-Nagel
Nucleobond PC100	Macherey-Nagel
Nucleobond PC100 EF	Macherey-Nagel
NucleoSpin Gel and PCR Clean-up	Macherey-Nagel
AllPrep DNA/RNA/Protein	Qiagen
SuperScript III First-Strand Synthesis System	Thermo Fischer Scientific
BigDye Terminator v3.1 Sequencing	Applied Biosystems
QIAquick PCR Purification	Qiagen
Vectastain IgG Mouse ABC	Biozol
Peroxidase Substrate Kit DAB SK-4100	Vector Laboratories
Duolink In Situ Red Starter Kit Mouse/Rabbit	Sigma-Aldrich
Gibson Assembly MasterMix	New England Biolabs

3.1.8 Vectors

Table 3.7 Vectors

Name	Supplier
pSLAX13-5'FLAG	Dr. Jochen Hecht (Berlin)
pSLAX13-5'HA	Dr. Jochen Hecht (Berlin)
RCASBP(A)	Dr. Jochen Hecht (Berlin)
RCASBP(B)	Dr. Jochen Hecht (Berlin)

3.1.9 Bacterial Strains

Cloning steps were performed with an in-house E. coli Top10 (A.C. Stiege) strain.

3.1.10 Cell Culture Lines

DF1 cell line originating from chicken fibroblast was obtained from ATCC.

3.1.11 Animals

Fertilized eggs (Clean Eggs quality) for chicken micromass cultures were purchased from VALO BioMedia GmbH (Osterholz-Scharmbeck).

3.1.12 Instruments

Table 3.8 Instruments

Instrument	Model No. / Type	Supplier
Table Top centrifuge	5414D	Eppendorf
Chilling centrifuge	5417R	Eppendorf
Microtiter plate centrifuge	5416	Eppendorf
Chilling centrifuge	Avanti J-E	Beckman-Coulter
Rotor	JLA16250	Beckman-Coulter
Ultracentrifuge	L7-55	Beckman
Ultracentrifuge Rotor	SW 32-Ti	Beckman
Thermocycler	GeneAmp PCR System 2700, 2720 and 9700	Applied Biosystems
Stereomicroscope	MZ7-5	Leica
Camera	Axiocam MRc5	Zeiss
Light source	KL1500 LCD	Leica
Software	Axiovision 4.x	Zeiss
Confocal microscope	LSM700	Zeiss
Sequencer	Genome Analyzer IIX	Illumina
Cluster Station		Illumina
Sonicator	BioRuptor NextGen UCD-300	Diagenode
Photometer	NanoDrop 2000	Thermo Scientific
Microplate Reader	Spectra Max 250	Molecular Devices

3.1.13 Software

General and General Biological Software

Cloning was performed *in silico* with CloneManager. Sequencing results were aligned and analyzed with DNA Star Sequman software. Reference sequences were taken from PubMed. UCSC Genome Browser was used for visualization of ChIP-seq and RNA-seq data. Digital pictures were edited using Carl Zeiss Axiovision 4.8.2 and CorelDRAW Graphics Suite. Figures were composed using Inkscape 0.48. The bibliography was managed using Mendeley.

Bioinformatics Software for Analysis of ChIP-seq and RNA-Seq

Table 3.9 Specialized Bioinformatics Software

Software	Task	Source
FastQC	Quality Control fastq-sequencing files	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
BWA Aligner	Aligning NGS-sequenced reads to reference genome	(Li and Durbin, 2009)
SAM-Tools	Handling of SAM-files	(Li *et al.*, 2009)
Centrimo	Central enrichment of motifs	(Bailey and MacHanick, 2012)
FIMO	Quantification of motifs in peaks	(Grant, Bailey and Noble, 2011)
TOMTOM	Identification of similar motifs in database	(Gupta *et al.*, 2007)
SPP	Quality Control of ChIP-enrichment	(Kharchenko, Tolstorukov and Park, 2008)
BED-Tools	Handling of BED-files	(Quinlan and Hall, 2010)
MACS2	Detection of ChIP-enriched regions	(Zhang *et al.*, 2008)
IDR	Reproducibility of ChIP-seq experiments	(Li *et al.*, 2011; Landt *et al.*, 2012)
seqMINER	Read distribution analysis of ChIP-seq datasets	(Ye *et al.*, 2011)
webPRANK	Sequence homology clustering	(Löytynoja and Goldman, 2010)
STAR mapper	Aligning NGS-sequenced reads to reference genome	(Dobin *et al.*, 2013)
Deseq2	Identification of differentially regulated genes	(Love, Huber and Anders, 2014)
GOrilla	Gene ontology	(Eden *et al.*, 2009)
GREAT	Gene ontology	(McLean *et al.*, 2010)
RStudio	Used for PCA and diff. reg. gene clustering	https://www.r-project.org/

3.2 Methods

3.2.1 Molecular Biological Methods

All standard molecular biological procedures were performed according to (Sambrook & Russell (2001).

3.2.1.1 DNA Isolation

Plasmid-DNA isolation from Top 10 competent cells was performed using the Nucleospin or Nucleobond PC100 kits (Macherey-Nagel), according to the manufacturer's instructions. Plasmid-DNA that was used for the cell transfection was always performed with Nucleobond PC100 EF kit.

3.2.1.2 RNA Isolation

RNA was isolated from the chMM cells using AllPrep DNA/RNA/Protein (Qiagen) according to manufacturer's instructions. First, chMM were harvested as for ChIP-seq, and separated just before fixation (approx. 300 000 cells from every sample) for protein and RNA extraction. Cells were spun down, washed with 1x DPBS, and 600 RLTμL buffer (supplemented with 0.001% β-mercaptoethanol) was added onto the cells. Extra RNase-Free DNase (Qiagen) treatment was added onto the columns to assure there is no DNA contamination, according to the manufacturer's instructions.

RNA from embryonic limbs at the HH25 was isolated the same way like RNA from the chMM cells with the additive step of tissue homogenization which was performed with a 0.5mm needle and a 1mL syringe.

3.2.1.3 Protein Isolation

Proteins were isolated from chMM cultures using the AllPrep DNA/RNA/Protein (Qiagen) kit according to manufacturer's instructions. See **Chapter 3.2.1.2** for the cell collection details.

3.2.1.4 Generation of cDNA

cDNA synthesis for qRT-PCR was performed with the SuperScript III (Thermo Fischer Scientific) using 1µg of total RNA as template and oligodT primers. The reaction was conducted in 20µl volume.

3.2.1.5 Polymerase Chain Reaction (PCR)

Standard PCRs were performed according to (Sambrook and Russell, 2001). PCR reactions were performed using in-house Taq and Pfu polymerases produced by A.C. Stiege for colony PCR, and Phusion DNA-polymerase (NEB) or Deep Vent DNA-polymerase (NEB) for cloning of the *HOX, HOXA10* deletion constructs, and *CTCF*. Additionally, PCR reaction of sequences with high GC content was supplemented with 1.5%-3% DMSO (Sigma-Aldrich).

PCR for HOXA10 deletion constructs was performed with long-range buffer (Roche, #11681842001), and specific reaction conditions indicated below.

Table 3.10 Reagents for the Amplification of *HOXA10* Deletion Constructs

Reagents	Amount
pSlax 3xFLAG-HOXA10	60ng
Primer F (10 pmol/µl)	2µL
Primer R (10 pmol/µl)	2µL
dNTPs (10mM)	1.5µL
DMSO	2.5µL
Long Range Buffer (before use incubate 10 min at 56°C)	5µL
ddH_2O	to 50µL

Table 3.11 PCR Machine Protocol for the Amplification of *HOXA10* Deletion Constructs

PCR machine protocol		
94°C	3´	
94°C	30´´	
65°C (-1°C per cycle)	1´	10x
68°C	9´(+30´´)	
94°C	30´´	
58°C	1´	20x
68°C	14´(+20´´)	
68°C	10´	
4°C	∞	

3.2.1.6 Sanger Sequencing

All PCR reactions for Sanger sequencing were performed with 80-100ng DNA as a template and BigDye v3.1 kit (Applied Biosystems), according to manufacturer's instructions. If DNA template had a high amount of the GC content, 1.5-3%, DMSO was added to the PCR reaction. The product was cleaned with ethanol precipitation before it was sent for sequencing, which was carried out by Mohsen Karbasiyan (Charité) on an ABI3700 capillary sequencer.

3.2.1.7 Cloning

CDS regions of the genes listed in **Table 3.12** were amplified with primers in **Table 3.3** for every gene individually (except *HOXD13*, see below) from HH25 limb bud cDNA using standard PCR and Pfusion polymerase (NEB). Amplification of *HOXA10* deletion constructs was carried out from a pSlax 5'-3xFLAG-*HOXA10* template, as explained in the **Chapter 3.2.1.5.**

Cloning of *HOX* Genes to pSLAX

pSlax vector was previously modified to contain 3xFLAG-tag and just after restriction sites for *Nco*I and *Spe*I (done by A.C. Stiege and J. Hecht). Therefore, the *Nco*I and *Spe*I were used as a cloning strategy for nine *HOX* genes. Forward amplification primers for *HOX* genes were designed so that they contain a restriction site for an enzyme creating an overhang identical to *Nco*I because *Nco*I sites were present in all *HOX* CDS (**Table 3.12**). Reverse primers were designed so they contain a site for *Spe*I. Amplified *HOX* gene products and pSlax 5'-3xFLAG were digested with *Nco*I (or an appropriate alternative) and *Spe*I, cleaned with NucleoSpin Gel and PCR Clean-up (Macherey-Nagel) and ligated overnight.

HOXD13 gene was especially difficult to verify by Sanger sequencing after amplification. Therefore a vector containing chicken *HOXD13* CDS was purchased from AddGene (CT#441), amplified with primers stated in **Table 3.3** and Deep Vent polymerase, and cloned into pSlax 5'-3xFLAG as described above.

Cloning of *CTCF* Gene to pSLAX

CTCF CDS region was amplified using primers designed for Gibson Assembly (**Table 3.3**). Then, the product was cleaned with NucleoSpin Gel and PCR Clean-up (Macherey-Nagel), and a Gibson Assembly was performed using Gibson Assembly MasterMix (NEB) with *CTCF* CDS and pSlax 5'-3xFLAG, or pSlax 5'-HA, following the manufacturer's instructions.

Cloning of *HOXA10* Deletion Constructs to pSlax

HOXA10 deletion constructs primers (**Table 3.4**) were designed to have 5‘ phosphorylated overhang and to point either forward or reverse from the end points of the desired deletion. Then, pSlax 5’-3xFLAG-*HOXA10* was used as a template for a long-range PCR as explained in the **Chapter 3.2.1.5**. Amplified pSlax 5’-3xFLAG-*HOXA10* deletion constructs were then simply ligated as they contained phosphorylated overhangs.

Cloning of *HOX*, *CTCF* and *HOXA10* Deletion Constructs to RCASBP

All constructs were cut from pSlax using *Cla*I and *Spe*I enzymes. This way tag sequence (3xFLAG or HA) was cut out together with the gene. At the same time, RCASBP(A) and RCASBP(B) were digested with *Cla*I and *Spe*I. Then, genes were ligated to the appropriate RCAS vector using standard overnight ligation.

Table 3.12 Reference Sequences

Sequence name	NCBI Reference Sequence	Enzymes used for cloning to pSlax
HOXA9	XM_003640734.1	*Esp3*I and *Spe*I
HOXA10	XM_001235692.1	*Esp3*I and *Spe*I
HOXA11	NM_204619.1	*BspH*I and *Spe*I
HOXA13	NM_204139.1	*Pci*I and *Spe*I
HOXD9	XM_001234506.2	*BspH*I and *Spe*I
HOXD10	XM_001234538.2	*Pci*I and *Spe*I
HOXD11	NM_204620.1	*BspH*I and *Spe*I
HOXD12	NM_205249.1	*Esp3*I and *Spe*I
HOXD13	As purchased in plasmid	*Bpi*I and *Spe*I
CTCF	NM_205332.4	Gibson Assembly cloning

3.3 Cell Culture Methods

3.3.1.1 Cultivation of DF1 cells

DF1 cells were thawed from the liquid N_2 stock, spun down to get rid of the DMSO and plated to 75cm^2 flask in the DF1 Standard. Cells were split 1:10 upon confluency and not kept in the culture past passage 50.

Table 3.13 DF1 Cultivating media

Media	Composition
DF1 Standard	DMEM, 4.5g/L Glucose; 10% FCS Superior (Biochrom, #S0615/0411A); 2% chicken serum (Gibco, #16110-082); 1% L-glutamine (Lonza, #17-605E); 1% penicilin-streptomycin (Biochrom, #A2213)
DF1 Starvation	DMEM, 1g/L Glucose; 1% FCS Superior (Biochrom, #S0615/0411A); 0.2% chicken serum (Gibco, #16110-082); 1% L-glutamine (Lonza, #17-605E); 1% penicilin-streptomycin (Biochrom, #A2213)

3.3.1.2 Virus Preparation

Virus preparation was performed as in Seemann (2006) with few modifications at the DF1 transfection step.

DF1 cells were grown on the two 10cm plates until confluency reached 80%. Transfection was performed in every plate as follows: 7µg of RCASBP DNA and 250µL 150mM NaCl were added to the first tube and mixed vigorously; 46.5µL of PEI (Polyscience, #24765-2) and 625µL of 150mM NaCl were added to a second tube and mixed vigorously. Then, contents of two tubes were combined, mixed vigorously, and incubated under the cell culture hood for 30 min. Thereupon, transfection mix was added to cells with a pipette.

3.3.1.3 Chicken Micromass (chMM)

Original chicken micromass protocol DeLise et al. (2000) was adapted by Seemann (2006).

Here, fresh, fertilized eggs were incubated for 4.5 days (to reach HH24-25 stage) at the 37.5°C and >60% humidity. After that, the eggs were opened and embryonic fore- and hindlimbs harvested and collected in PBS.

Then, limbs were washed 3-5 times with pre-warmed HBSS (Hanks' Balanced Salt Solution, Cambrex, #BE10-547F) solution. Limbs were digested with 300µL dispase solution (Gibco, #17105-041, 3 mg/ml in HBSS) for 15 minutes with gentle shaking every few minutes, to detach ectoderm. Limb buds were, then, washed 5-6 times with pre-warmed HBSS to ensure the near complete removal of the ectodermal layer. Then, limb buds were incubated for 30 min in 1mL pre-warmed digestion solution (0.1% (w/v) Collagenase type Ia (Sigma-Aldrich, #C9891), 0,1%

(w/v) trypsin (Gibco), 5% FBS in PBS) to detach the cells from one another. To ensure this 19mL of pre-warmed chMM media (DMEM Ham's F-12 (1:1) ; 1% FCS Superior (Biochrom, #S0615/0411A); 0.2% chicken serum (Gibco, #16110-082); 1% L-glutamine (Lonza, #17-605E); 1% penicillin-streptomycin (Biochrom, #A2213)) was addedto cells and passed through a cell strainer. Cells were then counted (Neubauer cell chamber), spun down, and adjusted in pre-warmed chMM media to a $2x10^7$ cells/mL cell suspension density.

Finally, cell suspension was split into aliquots, infected with an appropriate virus (virus-to-cell ratio 2 :1) and seeded in a 24-well plate in 10µL per well central droplets. Cells were left to attach to the plate for 2h at 37°C, and then 1mL of warm chMM media was carefully added to the cultures. chMM media was changed every two days and cultures were harvested for ChIP-seq, RNA-seq, and protein analysis six days post infection.

Staining of chMM Cultures

Individual chMM cultures were stained with Alcian Blue and Eosin every three days, until day 15, to assess differentiation process and impact of the overexpressing TFs onto this process.

For Alcian Blue, cultures were fixed for 1h at room temperature or 4°C overnight with Kahles Fixative (30%EtOH, 0.4%PFA, 4% acetic acid) and stained with After Eight Blue stain (0.05% Alcian Blue in 0.1M HCl-solution) overnight at room temperature. Then, cultures were washed with PBS, photographed with no liquid, left to dry, and destained overnight with 6M Guanidin Hydrochloride. Blue staining of chMM was then quantified on 592nm wavelength on ELISA plate reader.

For Eosin, cultures were fixed with 4% PFA in PBS at 4°C overnight. They were washed with PBS and stained with eosin for 60-90 sec. Cultures were photographed immediately after staining with no liquid in the well to prevent quick de-staining of Eosin.

3.4 Biochemical Methods

3.4.1 Determination of Protein Concentration

Protein concentration was measured using Bradford Assay (Roth) according to manufacturer's instructions. Absorption was measured at the 592nm wavelength on an ELISA plate reader. Proteins were then compared to known BSA dilution series and quantified.

3.4.2 SDS-PAGE

Protein gel electrophoresis (SDS_PAGE) was performed on 13.5% polyacrylamide gels as in (Sambrook and Russell, 2001)

3.4.3 Western Blot (WB)

Western blots were carried out according to standard procedures as in Sambrook & Russell (2001). Protein transfer was performed in a tank using PVDF membrane (Millipore Immobilon-P, 0.45µm pore size) and a pre-cooled transfer buffer (25mM Tris-Base; 200mM Glycine, 20% Methanol) at 4°C and 20V overnight.

Membrane was blocked with 3% BSA (Sigma-Aldrich) in TBS-T (0.1% Tween-20) for 1h. Then, primary antibody was incubated in the blocking solution at 1:1000 for 1h at room temperature. Membrane was washed with TBS-T three times for 10min at room temperature, and then secondary antibody was diluted in blocking solution at 1:1000 and incubated for 1h at room temperature. Finally, three TBS-T 10min washing steps were performed and the signal was detected using Western Lightning Plus-ECL (Perkin-Elmer) and various X-Ray films, developed on the AGFA Curix 60.

3.4.4 Co-Immunoprecipitation (co-IP)

Co-immunoprecipitation was performed following an adapted Andrews & Faller (1991) nuclear lysate co-IP protocol.

Cell Harvesting and Nuclear Extraction

First, DF1 cells were cotransfected with RCASBP(A)-5'-3xFLAG-*HOXA10*, RCASBP(A)-5'-3xFLAG-*HOXD13*, or any of seven RCASBP(A)-5'-3xFLAG-*HOXA10* deletion constructs and

RCASBP(B)-5'-HA-*CTCF* following standard transfection protocol as explained in **Chapter 3.3.1.2**. Cells were cultured with DF1 standard medium for six days, trypsinized (Gibco), resuspended in 1.5mL ice-cold DPBS and spun down for exactly 10sec on table top centrifuge. Supernatant was discarded, and resuspended in 400µL ice-cold Low-salt buffer (with freshly added Roche protease inhibitor) and incubated for 10min on ice. Cells were then vortexed for 10sec and briefly spun down on the table top centrifuge. Pellet was resuspended in 100µL ice-cold High-salt buffer (with freshly added Roche protease inhibitor) and incubated on ice for 20 min. Finally, nuclei were pelleted for 2min at 13 000 rpm at 4°C, and supernatant was transferred to a new tube and stored at -80°C.

Immunoprecipitation

25µL of magnetic Protein G (Invitrogen) beads were washed with 500µL of Blocking buffer three times. Afterward, 200µL of Blocking buffer was added to the blocked beads and, either, 5µL of αHA (BioLegends) or IgG control (Cell Signalling Technologies). The beads-antibody mix was incubated 3h at 4°C. Beads were washed three times with 300µL of BS3 Conjugation Buffer and then incubated with 250µL of BS3 Working solution at room temperature for 30 minutes on a rolling machine. The reaction was quenched with 12.5µL of Quenching buffer for 15 minutes at room temperature on a rolling machine. Beads were then washed two times with 500µL of cold IP buffer (with freshly added Roche protease inhibitor).

Next, the sample was added and diluted to 250µL with the IP buffer (with freshly added Roche protease inhibitor), and samples were incubated at 4°C overnight with rolling. Next day, the samples were washed two times with 1mL Wash buffer 105 (with freshly added Roche protease inhibitor) and three times with Wash buffer 150 (with freshly added Roche protease inhibitor) and then eluted with 30µL co-IP elution buffer for 15min at room temperature with rolling. Beads were then spun down at 13 000 rpm for 1min, and the supernatant was stored at -80°C.

Immunodetection

Eluted samples were then loaded on an SDS-PAGE gel, and subsequently, Western blot was performed using αFLAG antibody (Sigma-Aldrich) as described in the chapter **Chapter 3.4.2** and **Chapter 3.4.3**.

3.5 Proximity Ligation Assay (PLA)

PLA was performed using Duolink In Situ Red Starter Kit Mouse/Rabbit (Sigma-Aldrich) according to manufacturer's instructions with some in-house solutions.

First, DF1 cells were transfected as shown in **Chapter 3.3.1.2** and cultured for six days. Then, they were re-plated to 24-well plate with coverslips to optimize the surface area for the PLA. Next day, cells were fixed with 4% PFA in PBS for 11min and washed three times with 1x PBS.

Cells were blocked with an in-house PLA Blocking solution for 1h at room temperature. Then, primary antibodies αHA (BioLegends), αRAD21 (Abcam), αCTCF (Active Motif), and αFLAG (Sigma-Aldrich) were diluted with an in-house PLA antibody diluent and 200µL this dilution was added to cells. Antibody dilutions were then applied to cells (as in **Table 3.14**) and incubated for 1h at 4°C. Afterward; slides were washed two times 5min with 10% horse serum in TBS-T (0.2% Tween).

PLUS and MINUS probes (Duolink kit,secondary antibodies labeled with oligos) were diluted 1:5 with PLA antibody diluent and incubated at room temperature for 20min. Diluted PLUS and MINUS probes were added to cells and incubated in a pre-heated humidity chamber at 37°C for 1h. Cells were, then, washed two times 5min with 1x PLA buffer A.

Ligation stock (Duolink kit) was diluted 1:5 with ddH_2O and ligase was added last in 1:40 ratio. 40µ of the ligation reaction was added to cells and then incubated in a pre-heated humidity chamber at 37°C for 30min. Cells were then, washed two times 2min with 1 x PLA buffer A.

Then, polymerase chain reaction was performed with fluorescently labelled nucleotides. First, Amplification stock (Duolink kit) was diluted 1:5 in ddH_2O and polymerase was added last in 1:80 ratio. Then, 40µL of diluted Amplification solution was applied to the cells and incubated for 100min in the preheated humidity chamber at 37°C. Cells were washed two times 10min with 1x PLA buffer B and once 1min with 0.01x PLA buffer B. All reactions from Amplification on were performed in the dark.

Table 3.14 PLA Experimental Setup, Antibodies, and Antibody Dilutions

	Transfection combinations	Desired detection	Antibodies used and dilutions
Positive control	RCASBP(B)-5'-HA-*CTCF*	CTCF-RAD21	rb αRAD21 (Abcam) 1:1 000 m αHA (BioLegends) 1:8 000
De-novo interactions	RCASBP(A)-5'-3xFLAG-*HOXA10*	HOXA10-CTCF	m αFLAG (Sigma-Aldrich) 1:20 000 rb αCTCF (Active Motif) 1:20 000
	RCASBP(A)-5'-3xFLAG-*HOXD13*	HOXD13-CTCF	m αFLAG (Sigma-Aldrich) 1:20 000 rb αCTCF (Active Motif) 1:20 000
	RCASBP(A)-5'-3xFLAG-*HOXA10* deletion constructs	HOXA10Δ-CTCF	m αFLAG (Sigma-Aldrich) 1:20 000 rb αCTCF (Active Motif) 1:20 000
Negative control			m αFLAG (Sigma-Aldrich) 1:20 000 rb αCTCF (Active Motif) 1:20 000

3.6 Chromatin Immunoprecipitation

Chromatin Immunoprecipitation (ChIP) is a technique that allows for identification of protein-DNA interactions. It is a week-long, multi-step protocol where a specific antibody is directed against a TFs, histones, histone modifications, or any other protein binding to DNA (e.g. RNA Pol II). Results of chromatin immunoprecipitation can be analysed on qRT-PCR or by next generation sequencing (NGS), which offers a genome-wide profile for protein binding. Here, ChIP was accompanied with NGS sequencing to study nine paralogous HOX-TFs, to determine and compare their DNA binding.

3.6.1 Chromatin Preparation

Here used ChIP protocol is a Lee et al. (2006) ChIP protocol modified by Ibrahim (2014) and subsequently by I.Jerković.

Cell Harvesting and Fixation

After six days of culture, chMM cultures were taken out of the culture, and the media was aspirated. Then, 200µL of collagenase solution (0.1% (w/v) Collagenase type Ia (Sigma-Aldrich, Cat.-No. C9891) in chMM-medium) was added to every well of the 24-well plate and incubated at 37°C for 30-90min. This incubation was followed by disruption by pipetting which allowed easier

detachment of chMM cultures from wells and easier disruption of a very dense clump of cells building the chMM culture. Harvested cultures were collected in a Falcon tube, pelleted for 5min at 1 000 rpm at 4°C, and pellet was resuspended in 10mL of cold chMM medium. Usually, 3-4 plates were used per replicate.

Cells were then crosslinked with 273.5µL of 37% formaldehyde (Roth) (1% final concentration) and incubated for 10min on ice with occasional turning the tube. As this is an important step, the fixation time was always rigorously controlled to achieve better comparability between the samples. If cells get fixed too strong it might induce a GC bias in the NGS, whereas if they are under fixed there will be many positive interactions lost as they won't be cross-linked efficiently. Therefore, after exactly 10min 550µL of 2.5M glycine in PBS (0.125M final concentration) was added and the tubes and inverted several times to quench the formaldehyde efficiently. After that, cells were washed twice with 1x PBS and cells were snap-frozen in liquid N_2 and stored at -80°C.

Cell Lysis and Nuclear Extraction

Pelleted cells were taken out of the -80°C, thawed on ice, resuspended in 10mL of ice-cold Lysis buffer 1 (with freshly added Roche protease inhibitor), and incubated for 10min at 4°C, with rocking. Then, cells were spun down (5min at 2700xg) at 4°C and supernatant discarded. Pellet was resuspended in 10mL of Lysis buffer 2 (with freshly added Roche protease inhibitor) followed by 10min incubation at room temperature with rocking. Cells (at this point already nuclei) were spun down again (5min at 2700xg), the supernatant discarded, and the pellet was resuspended in 1.5mL of ice-cold Lysis buffer 3 (with freshly added Roche protease inhibitor).

Sonication

Then, cells/nuclei were sonicated on the Bioruptor NextGen for 35 to 45 cycles (30seconds pulse, 30seconds pause) making sure that the water bath was cool entire time of sonication duration to prevent overheating of the samples.

Finally, samples were cleaned up from cellular debris by the addition of 10% Triton-X-100 (to 1% end concentration). Samples were, then, centrifugated at 16 000xg for 15min at 4°C and supernatant was transferred to a new tube. 5-10% of the sample volume was taken for the sonication quality control, and rest of the sample was stored at -80°C.

Sonication Control

Sample aliquot separated after sonication was used to assess quality of the sonicated chromatin. For this, 4µL of Proteinase K (20mg/mL) and 5M NaCl in final concentration of 10% was added to sonicated chromatin. Everything was then mixed and incubated at 65°C overnight. Next day, 4µL of RNaseA (5mg/mL) was added to de-crosslinked chromatin and incubated 30min at 37°C. DNA was, then, purified by ethanol precipitation and dissolved in 20µL ddH_2O. After that, DNA concentration was measured on Nanodrop to determine chromatin concentration, and the rest 19µL were loaded on a 1% agarose gel to investigate sonication efficiency (successful sonication smear appears between 200-500 bp size).

Chromatin concentration was determined using the following formula:

$$c\begin{pmatrix}chromatin\\ ng/\mu L\end{pmatrix} = \frac{c\begin{pmatrix}Chromatin\ for\ sonication\ control\\ as\ measured\ on\ nanodrop\end{pmatrix} * V\begin{pmatrix}Chromatin\ for\\ sonication\ control\end{pmatrix}}{V(total\ chromatin\ after\ sonication)}$$

3.6.2 Immunoprecipitation

Beads-antibody Preparation

First, 35µL of Protein G beads were washed three times with 1mL of 0.5% BSA in DPBS to block the beads. Then, 8µL of the αFLAG (Sigma-Aldrich), 8µL of αRAD21 (Abcam), 2.5µL of αH3K4me3 (Millipore), or 2.5µL of αH3K27me3 (Millipore) was diluted in 300µL of 0.5% BSA-DPBS. Antibody dilutions were added to blocked beads and incubated overnight at 4°C with rotation.

Immunoprecipitation

Next day, beads were washed with ice-cold Lysis buffer 3 (with freshly added Roche protease inhibitor), chromatin was added to beads-antibody mix, and volume was adjusted with Lysis buffer (with freshly added Roche protease inhibitor and Triton-X-100 (final concentration 1%)), to 1.2mL. If the ChIP was performed for TFs 35µg of chromatin was added to immunoprecipitation and if ChIP was performed for histone modifications 10-15µg of chromatin was used for immunoprecipitation. Samples were incubated overnight at 4°C with rotation. In parallel, 100µL of the non-IPed sonicated chromatin was taken as an input control and stored at -20°C.

Following day, samples were washed seven times using ice-cold RIPA buffer (with freshly added Roche protease inhibitor) and once with ice-cold TE-NaCl (with freshly added Roche protease inhibitor) buffer on magnetic rack. Supernatant was discarded, and beads were centrifuged at 1 000rpm at 4°C for 3min, and rest of the supernatant was carefully removed using the magnetic rack to prevent any carryover of beads.

Elution, Crosslink-reversal, and Purification

Proteins were then eluted by adding the 210µL of Elution buffer to beads and incubating them at 65°C for 30min with vigorous shaking. Afterward, beads were spun down at 16 000xg for 1min at 4°C, and 200µL of the supernatant was transferred to a new tube.

Then, the input chromatin was thawed and together with the ChIP'ed samples subjected to crosslink-reversal. For crosslink-reversal, first 4µL of Proteinase K (20mg/mL) and 5m NaCl to 10% end concentration were added to each sample, mixed, and incubated at 65°C overnight. Next day, 4µl of RNaseA (5mg/mL) was added to each sample and incubated at 37°C for 30min.

Finally, samples were purified using the standard Phenol-Chloroform method and precipitated with ethanol using 10% 3M Na-acetate and 4µL of glycogen (Ambion 5mg/ml, # AM9510) as a carrier. Precipitated samples were sent to the BCRT sequencing facility for further processing. Input and ChIP samples were sequenced on Illumina HiSeq in 50bp single-end reads.

3.6.3 Quality Control and Initial Processing of ChIP-seq data

ChIP performed in this study was used in combination with NGS to create genome-wide binding HOX maps. Since the ChIP-seq initial output generates millions of sequencing reads, it was necessary to perform comprehensive and complete quality control on individual samples, and reproducibility controls on biological replicates. Sequenced reads were enriched for fragments that come from the immunoprecipitated DNA, therefore allowing for the visualization of these enrichment points, called peaks. In order to determine if sample was of good quality, peak-to-background ration had to be high, and to determine reproducibility between samples peaks had to always occur at the same positions. Roughly, these two issues are focal points of ChIP-seq quality analysis and will be discussed in more detail in subsequent chapters. Initial control analysis and reproducibility testing were performed according to the ENCODE guidelines for ChIP-seq (Li *et al.*, 2011; Landt *et al.*, 2012).

ENCODE guidelines were followed closely, and all the necessary tools to run this pipeline were available on the locally installed GALAXY platform at the Charitè (Blankenberg *et al.*, 2010; Goecks *et al.*, 2010). This local GALAXY platform was installed and moderated by Peter Hansen, and the ENCODE pipeline was designed and implemented by Peter Hansen and Daniel M Ibrahim.

Quality Control and Read Mapping

First step in the assessment of a ChIP-seq sample is the quality control of raw sequences. This was assessed with the FastQC program (Filter-by-Quality function (FASTX-toolkit)) which allowed all the reads that did not score well enough (if average Phred-score < 28) to be discarded. Rest of the sequences were then mapped to chicken genome (galGal4) with the BWA aligner using following parameters (aln -n=0, aln -o=1, aln -e=1, aln -d=16, aln -i=5, aln -l=-1, aln -k=2, aln -M=3, aln -O=11, aln -E=4, aln -R=FALSE, aln –N=FALSE, samse/sampe –n=3, sampe=–N10, sampe –a=500, sampe –o=100000, samse/sampe –r=NO). BWA aligner created an output .sam file with mapped reads. .sam file was, then, used to remove the reads that map ambiguously in the genome (e.g. repetitive regions) by selecting lines matching XT:A:U, @SQ, or @PG (XT:A:U- a .sam tag that marks uniquely mapped reads, @SG, and @PG are comment lines). Then, a .bam file was created and PCR artifacts (introduces during the library preparation) were removed with rmdup function (SAMtools).

Once all poor quality, ambiguous, and redundant reads were kicked out, filtered .bam file was subjected to quality of enrichment analysis. This was assessed using cross-correlation analysis of the spp 1.11 package (get-binding characteristics function, parameters: -srange=0,1000; -bin=5; -cluster=2; -debug=F; -MinTagCount=1000; -Acceptance_Z_Score=3; -RemoveTagAnomalies=T; -Anomalies_Z=5; -AcceptAllTags=F).

ENCODE guidelines suggest different quality thresholds, however not all thresholds have to be met by every experiment. Most relevant thresholds are the sequencing depth, NRF, NSC, and RSC. Sequencing depth was aimed to reach at least 10 000 000 non-redundant reads per biological replicate. This enables the recovery of almost all binding sites in the genome. NRF (non-redundancy fraction) compares the ratio of redundant (PCR artifact) and non-redundant reads, giving information about the library complexity. NSC (normalized strand coefficient) and RSC(relative strand coefficient) and two parameters that provide an approximation of enrichment in the ChIP. However, these two metrics can also be influenced by the biological properties of

investigated TFs. Some ChIP-seq data scored marginally on this analysis. This is further discussed in **Chapter 4.5.1** and **Chapter 5.2.1**.

Reproducibility Analysis

After initial and quality control of individual samples, it is essential to address reproducibility between the samples. This was performed according to the ENCODE guidelines with the Irreproducibility Discovery Rate analysis (IDR) (Li *et al.*, 2011).

IDR is used for two main purposes. First, IDR analysis compares similarity between two samples. It takes two ranked lists of peaks (containing both true peaks and false peaks) and compares them. IDR assumes that the higher the peak is, the more reproducible it should be. By using true positive and false positive the IDR can discriminate between groups.Second, IDR analysis helps determine the threshold for the reproducible peaks.

Peak Calling

In order to call reproducible peaks from a ChIP-seq experiment, an IDR analysis has to be run first. For this, it is necessary to call peaks with low stringency. This is done with MACS2 (parameters: -g 1.0e9, –p 1e-1, --too-large, --bw (as determined by spp in the quality control step)). MACS2 calls a huge number of peaks (300 000 or more) most of which are not true peaks. For IDR analysis, peak lists are truncated to 120 000 and ranked according to the p-value to inform of the peak strength. To compare the reproducibility between two biological replicates peaks have to be called for eight datasets. Peaks are called on two biological replicates to check for the *replicate consistency*. Furthermore, peaks are called for pseudoreplicates[2] for each of the two biological replicates to check for *self-consistency*, and on a pooled sample to check *pooled-sample reproducibility*.

Self-consistency (pseudoreplicates)	Rreplicate consistency	Pooled consistency
Rep1.1 Rep1.2	Rep1	Rep0.1
Rep2.1 Rep2.2	Rep2	Rep0.2

[2] Pseudoreplicates are randomly selected reads from the parent .bam file (e.g. Rep1 .bam reads are randomly selected and distributed to two new .bam files creating pseudoreplicates Rep1.1 and Rep1.2).

Irreproducibility Discovery Rate (IDR)

IDR was performed for every of the eight peak lists effectively checking the *self-consistency*, *replicate consistency*, and *pooled-consistency*.

IDR informs which number of peaks are reproducible above a certain threshold. For example, a number of peaks determined for the pooled-sample above the 0.01 threshold means that 99% of the peaks under that threshold is reproducible. For *replicate consistency* and *self-consistency* IDR thresholds of 0.01 were used for *pooled-consistency* 0.0025, except for the HOXA13. For the HOXA13 we used a "rescue strategy", as suggested by ENCODE, using reproducible peak number from 0.01 threshold from *pooled-consistency*. This was chosen because second biological replicate of HOXA13ChIP-seq had same binding profile like the first biological replicate, but lower signal-to-noise ratio.

Final Peak Sets

To determine final peak sets, peaks were called using MACS2 with low stringency. This set of peaks was then sorted on the p-value. From this ranked list, peak number was determined by the *replicate consistency* 0.01 threshold, and top confidence peaks were selected. This was adjusted to *pooled-consistency* 0.01 threshold for the HOXA13 peaks.

3.7 Bioinformatics Analyses

3.7.1 Motif Analysis

3.7.1.1 *De novo* motif analysis

Peak summits were extended +/-75bp around the summit using bedtools (getfasta -name -s –fi galGal4genome.fa -bed input.bed -fo output.fa). Extended fasta files were used as an input for RSAT tool (Thomas-Chollier *et al.*, 2012) where fast and efficient, *de-novo* motif analysis was performed using default parameters. Oligomer length selected was set to 7.

3.7.1.2 Peak Overlap & Peaks in Regions

Peak Overlaps

Peak summit .bed files were extended +/-150bp around the summit. Then bedtools (intersect -a A.bed -b B.bed -f 0.33222591362126245847176079734219 -u -wa >AoverlapB.bed) were used to determine overlaps between two peaks. Two peak summits had to be no further than 200bp apart to be considered as overlapping.

Overlaps were performed in the following manner. First, top 10 000 peaks from the dataset A were investigated for overlaps with entire dataset B. Then, top 10 000 peaks from dataset B were investigated for overlaps with entire dataset A (see **Chapter 4.5.2**). This was done to accomplish better comparability between the samples.

Peaks in Regions

galGal4 genome was partitioned into "regions": introns, exons, promoters (-5kb up to 2kb around TSS), and intergenic (the rest of the genome not attributed by aforementioned three) regions. Then overlaps between peaks and "regions" were obtained as described in Peak Overlaps above.

3.7.1.3 MEME Analyses

Centrimo

Peak summits were extended +/-250bp around the summit and .fasta file was produced, as explained in **Chapter 3.7.1.1**. Centrimo analysis was performed using default parameters (Bailey and MacHanick, 2012).

FIMO

Peak summits were extended +/-150bp around the summit and .fasta file was produced, as explained in **Chapter 3.7.1.1**. Peaks were then screened for the individual uploaded motifs using default parameters (Grant, Bailey and Noble, 2011). For the motif count, it was made sure peaks were counted only once if carrying more than one motif.

TOMTOM

TOMTOM was used for the comparison of the A9-PBX1 compact motif, AP1, "Unmatched", and CTCF motif comparisons using default parameters (Gupta *et al.*, 2007).

3.7.2 seqMINER

seqMINER was used to generate overlaps between peaks (HOX and CTCF) and genome-wide binding profiles (histones). Parameters were set as follows, 5 clusters, KMeans seed 3150032, +/-5kb window, and KMeans linear clustering normalization for multiple datasets (Ye *et al.*, 2011).

3.7.3 Principal Component Analysis (PCA)

galGal4 genome was partitioned in 500bp windows and then overlapped with the peak summits. This created a binary table that was made into a matrix (with mat_data). PCA was performed using FactoMineR R package.

3.7.4 Vista Enhancer Clustering

Vista enhancer clustering was performed using same logic like in the **Chapter 3.7.3**. Binding was examined in every enhancer region, and a binary table was created. Later this table was made into a matrix (with mat_data) and clustered in R using heatmap2.

3.7.5 RNA-seq

RNA-Seq mapping and differential gene expression were performed using a pipeline created for RG Mundlos by Dr. Stefan Haas.

Mapping and Annotation of RNA-Seq data (with the S.Haas pipeline)

RNA-seq reads were mapped to galGal4 genome using STAR mapper (Dobin *et al.*, 2013). Splice junctions were based on RefSeq/ENSEMBL combined gene annotations options included: alignIntronMin 20, alignIntronMax 500000, outFilterMultimapNmax 5, outFilterMismatchNmax 10, and—outFilterMismatchNoverLmax 0.1). RefSeq (galGa4) and ENSEMBL (release 75) gene annotations were combined to generate read counts for the individual gene..

Generation of Differentially Regulated Genes (with S. Haas pipeline)

Differential expression of genes was generated using DEseq2 comparing either a *HOX* overexpression and control or two *HOX* paralogue count files (Love, Huber and Anders, 2014). Top 50 differentially regulated genes were identified with DEseq2 (top-value $< 10^{-5}$,minimum base mean >30, and a fold change >2) and clustered with R heatmap3 (log2 transformed changes (compared to control) were used as R input).

Gene Ontology Analysis

For GOrilla analysis gene names were used as input. Gene names were lifted over to mm9 gene names using Orthoretriever (UCSF, 2015). For a background all mm9 genes were used. GOrilla was used with default parameters (Eden *et al.*, 2009).

For GREAT analysis only genomic coordinates can be used. Genomic locations were retrieved using Orthoretriever and isolating the genomic coordinated from the output. GREAT was used with default parameters(McLean *et al.*, 2010).

Estimation of Viral Expression Levels Dr. Stefan Haas and Dr. Daniel M Ibrahim

RCAS codes for a polycistronic mRNA where only one-third of all transcripts contributes to the production of a HOX protein. To address this issue, all splice variants that could be unambiguously identified were counted and a ratio ("HOX factor") of the *HOX* splice variant was produced for every *HOX*. Then, *HOX* RPKM values were multiplied by its own "HOX factor" producing the real RPKM values.

In parallel, RPKM for all nine *HOX* genes from HH23 posterior, distal forelimb was calculated. In this tissue, nine *HOX* are at least partially expressed (see **Chapter 4.3**). Then, the chMM RPKM was simply divided by the forelimb RPKM, and the overexpression fold was generated.

4 Results

4.1 Investigation of Posterior HOXA/D Protein Homology

Hox genes have emerged from a single ancestor through series of tandem and whole genome duplications. As *Hox* gene number increased, it was hypothesized that the evolutionary pressure temporarily reduced creating a "window of evolvability" (Wagner, Amemiya and Ruddle, 2003). These newly duplicated genes could then adopt new role and expression during embryogenesis and allow for the morphogenesis of new organs and structures. This novel role would emerge through the change in the *cis*-regulatory elements or the changes in the gene body. The changes in the gene body inherently affect its protein product and the protein function, possibly affecting the Homeodomain and the protein-DNA binding.

Therefore, I first examined how similar are HOX protein primary sequences and if differences in protein content mirror the differences in protein-DNA binding. For this, protein sequences were

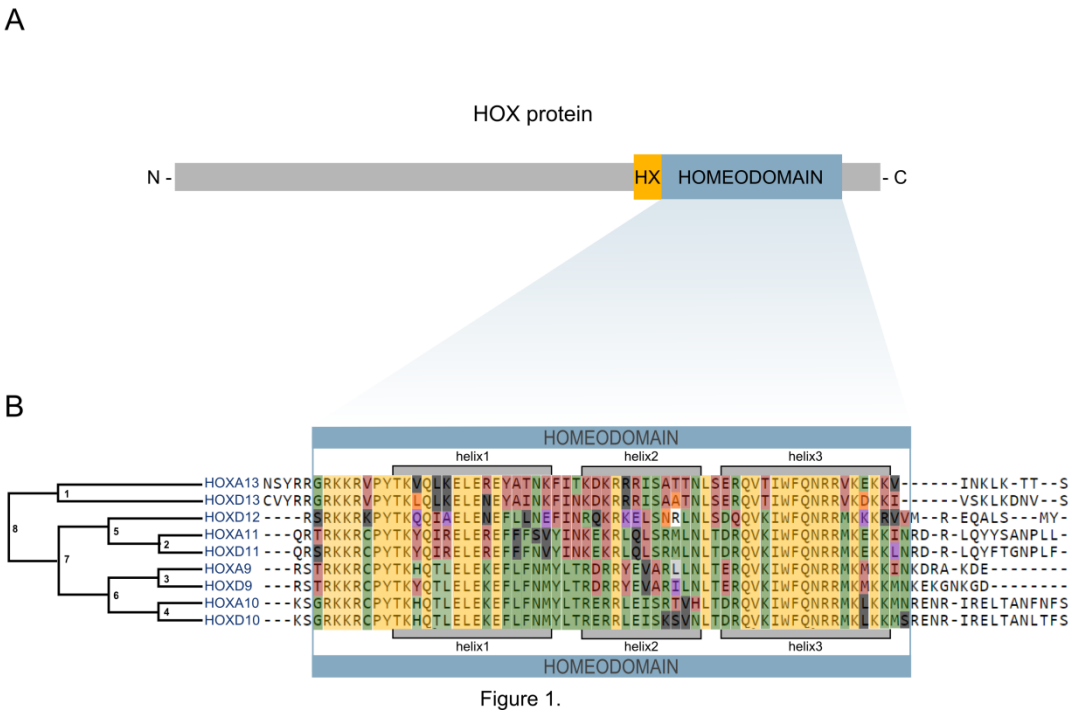


Figure 1.

Figure 4.1 Posterior HOX protein sequence recapitulates the evolutionary homology.
A) Schematic drawing of a HOX protein with a yellow box indicating the position of a Hexapeptide (HX) and a blue box indicating a position of the DNA binding domain, Homeodomain. B) Clustering and dendrogram based on full HOXA9-13 and HOXD9-13 protein sequences. Clustering was performed with webPRANK, European Bioinformatics Institute with default settings (Löytynoja and Goldman, 2010).

extracted from PubMed,compiled, and used as an input for the webPRANK software (Löytynoja and Goldman, 2010). Sequences were clustered according to protein similarity. Additionally, the C-terminal part of proteins, carrying the Homeodomain, was annotated for easier comparison (**Figure 4.1A** and **B**). Expectedly, this analysis confirmed Homeodomain is the most conserved part of the HOX protein. Furthermore, individual paralogy group (PG) TFs (e.g. PG13=HOXA13 & HOXD13 or PG11=HOXA11 & HOXD11) always clustered closest to each other (Garcia-Fernandez, 2005).

The analysis also demonstrated that HOX13 TFs diverged somewhat further apart from the rest of the HOX-TFs, indicating that HOX13 protein ancestor probably acquired some changes before diverging into HOXA13 and HOXD13. Surprisingly, although HOXD12 TF is functionally redundant with HOXD13, it clustered closer with more anterior HOX proteins (here, HOX9 and HOX10) than with HOX13 indicating that protein sequence alone is not a suitable functional indicator. Finally, I examined protein conservation of the three Homeodomain (HD) helices. For this, the three alpha helices were annotated in their respective Homeodomain and then examined for the conservation (**Figure 4.1B**, in yellow). Expectedly, helix 3 was the most conserved amongst nine HOX-TFs, as most of the sequence recognition and the protein-DNA contact is established through this helix.

4.2 Characterization of chMM as a System for Functional HOX Investigation

Transcription factors are nuclear proteins that bind motifs on the DNA and control the transcriptional output of their target genes. To investigate a transcription factor binding and, its regulatory and functional effects, one has to have a highly specific antibody and a sufficient amount of tissue available. In this study, however, to satisfy these two simple criteria it was necessary to design a novel approach.

Firstly, to investigate the binding of nine posterior HOX-TFs, it was necessary to discriminate between the individual proteins. As demonstrated by the conservation analysis, HOX proteins are remarkably homologous, particularly in their DNA binding domain (**Figure 4.1B**). Therefore, there is no available antibody, commercial or homemade, that can distinguish between these nine HOX proteins. Furthermore, overlapping pattern expression prevented the use of a *meta*-antibody. To overcome these issues, nine posterior *HOX* genes (*HOXA9*, *A10*, *A11*, *A13* and

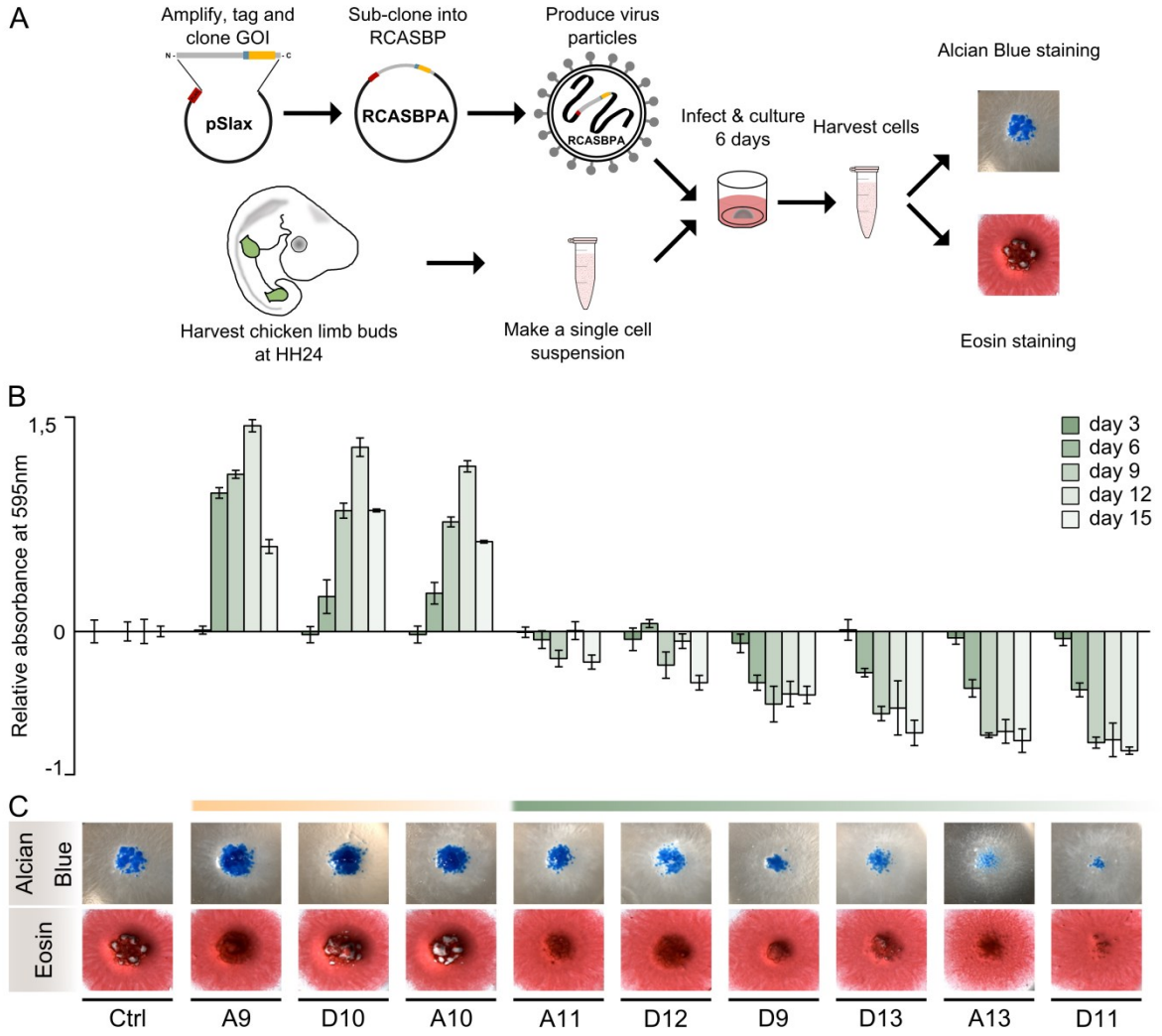


Figure 4.2 Over-expression of the HOX-TFs induces specific transcriptional programs that shape chondrogenesis in the chMM.
A) Alcian Blue quantification of the chMM cultures that express individual HOX-TFs, at the five time points, day 3, 6, 9, 12 and 15 post infection. B) Alcian Blue and Eosin staining of the chMM cultures, here visualized at day 9.

HOXD9, *D10*, *D11*, *D12,* and *D13)* were FLAG-tagged and cloned into an RCASBP(A) vector (**Figure 4.2A**). In this way, I introduced a unique epitope for which a highly specific antibody was readily available.

Secondly, to obtain sufficient amount of the input material chicken micromass (chMM) system was adapted. In this system, mesenchymal stem cells (MSC) are harvested from chicken embryos at the HH24, infected with the RCASBP(A) virus and cultured for up to 15 days (**Figure 4.2A**) (Morgan and Fekete, 1996). chMM system is a chondrogenic cell culture system where isolated MSCs are plated in a high-density droplet ensuring hypoxic conditions that resemble conditions

in the developing limb bud (Morgan and Fekete, 1996). Importantly, at the HH24 stage, all of the investigated *HOX* genes are expressed at least in a subset of cells. Therefore, both, adequate cellular environment and HOX cofactors are present which is a prerequisite for a comprehensive investigation of the HOX-TF binding.

Thirdly, RCASBP(A) is a modified chicken retrovirus where the oncogene is replaced with the individual gene of interest (GOI) and moderately overexpressed (see **Chapter 4.3**. for details).

With these modifications, it is possible to obtain GOI with the unique epitope in a native cellular environment and study the binding of HOX-TFs as accurately as possible.

In this system, I first investigated the impact of overexpressed *HOX* genes to the chondrogenic process undergoing in the chMM cultures. This was surveyed with two histological staining: Eosin, to examine general cell morphology; and Alcian Blue, to determine the chondrogenic potential. Four individual cultures were fixed and stained on day 3, 6, 9, 12 and 15 post RCASBP(A)-*HOX*[3] infection. Stained cultures were, examined visually and compared to the control. Additionally, Alcian Blue staining was measured which allowed quantification of the chondrogenic potential in these cultures. Both, the appearance and quantification of the chMM cultures demonstrated that *HOX* genes can both, inhibit or accentuate the chondrogenic potential of the culture (**Figure 4.2B and C**). Interestingly, the extent of the inhibition or accentuation of chondrogenesis differed between the nine *HOX-TF*s, with no correlation to protein sequence conservation. However, the paralogy groups generally exhibited similar impact on the chondrogenesis, except PG9 (**Figure 4.2B** and **C**). More specifically, PG10 accentuates, whereas PG11, PG13, and HOXD12 inhibit chondrogenesis which is in line with known HOX-TF effects on the chondrogenesis (Kuss *et al.*, 2009; Ibrahim *et al.*, 2013). Surprisingly, however, HOXA9 and HOXD9 overexpression affect the cells oppositely, one inhibiting and the other exacerbating the undergoing chondrogenesis (**Figure 4.2B** and **C**).

On the other hand, Eosin staining gives information about the morphology of central chondrogenic condensation and the surrounding fibroblastic-like cells. Cultures stained with Eosin exhibited changes at both, chondrogenic and fibroblastic-like cells. Surrounding fibroblastic-like cells in all, but HOXA13 culture, appear in thick rays radiating from the central condensation. In the HOXA13 culture, these cells appear less dense and also populate smaller area than in other chMM cultures. These observations are consistent with the diverse functions *HOX* genes play in chondrogenesis (Raines *et al.*, 2015).

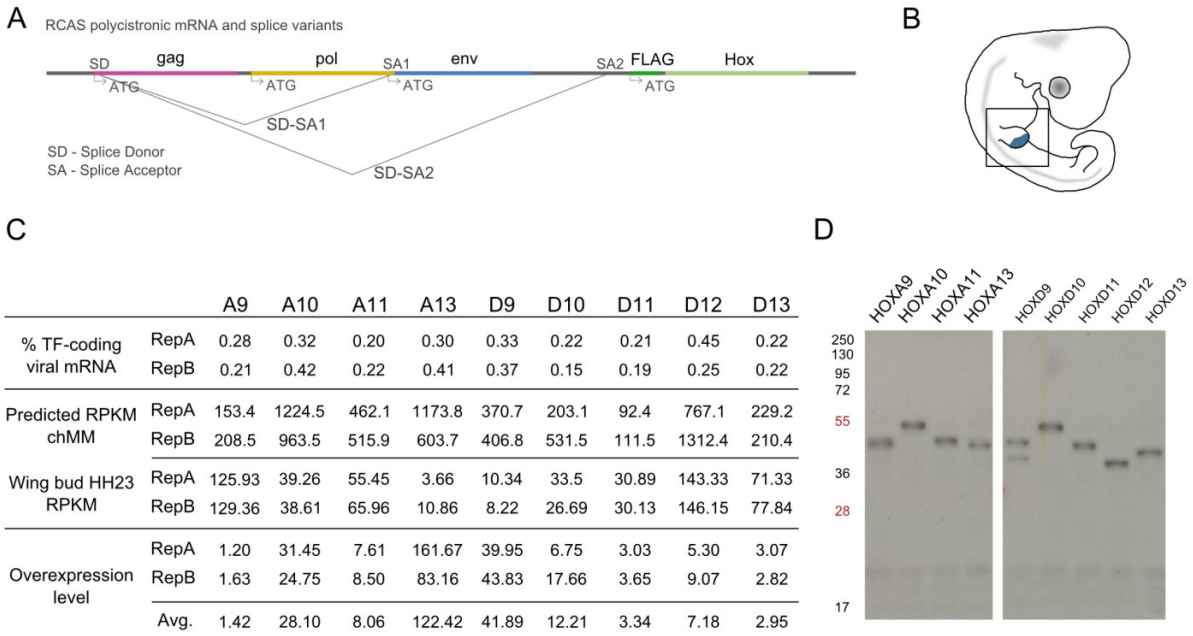
[3] chMM cultures were infected individually with nine *HOX* constructs.

Taken together, these preliminary analyses demonstrated that all HOX-TFs distinctly impact chondrogenic potential of the chMM culture and that paralogy groups often similarly affect undergoing chondrogenesis. Furthermore, these results corroborated previous findings and clarified effects of individual HOX-TFs to the process of chondrogenesis. Finally, these results confirm chMM is a suitable system for the further functional analyses of the HOX.

4.3 Estimation of Viral *HOX* gene Overexpression Levels

To ensure that the chosen methodology reaches all desired standards it was necessary to examine the abundance of overexpressed transcripts and proteins. Before addressing these issues, first, it is important to understand the here used system better. RCASBP(A) is a modified retrovirus that integrates into the genome and soon after the viral genes (and *HOX* genes) get expressed. Viral mRNA is polycistronic and splices in three different isoforms. Out of these three, only one splice variant will lead to the HOX protein product, causing quite moderate overexpression.

To compare *HOX* gene expression levels between the chMM system and an *in vivo* system, it was necessary to select appropriate chick embryo tissue subsection and stage. For this, posterior, distal chick forelimb, at the stage HH23, was dissected in two biological replicates (**Figure 4.3B**). Posterior, distal forelimb was selected as it was the subsection best representing the expression of nine *HOX* genes. Then, total RNA was extracted and processed for RNA-seq and comparison with the chMM overexpressions. Therefore, only one tissue was selected for the comparative expression analysis of nine genes. Due to complicated HOX expression patterns in the limb bud, this approach had its shortcomings. Namely: 1) *HOXA10* expression was present throughout the dissection but weakly, 2) only about 20% of the cells expressed the HOXA13, and 3) HOXD9 was expressed only in a subset of cells (Nelson *et al.*, 1996). Additionally, *HOXD11* gene was expressed in all dissected cells (Nelson *et al.*, 1996). This is relevant information, as it greatly facilitated correct estimation of the overexpression.



		A9	A10	A11	A13	D9	D10	D11	D12	D13
% TF-coding viral mRNA	RepA	0.28	0.32	0.20	0.30	0.33	0.22	0.21	0.45	0.22
	RepB	0.21	0.42	0.22	0.41	0.37	0.15	0.19	0.25	0.22
Predicted RPKM chMM	RepA	153.4	1224.5	462.1	1173.8	370.7	203.1	92.4	767.1	229.2
	RepB	208.5	963.5	515.9	603.7	406.8	531.5	111.5	1312.4	210.4
Wing bud HH23 RPKM	RepA	125.93	39.26	55.45	3.66	10.34	33.5	30.89	143.33	71.33
	RepB	129.36	38.61	65.96	10.86	8.22	26.69	30.13	146.15	77.84
Overexpression level	RepA	1.20	31.45	7.61	161.67	39.95	6.75	3.03	5.30	3.07
	RepB	1.63	24.75	8.50	83.16	43.83	17.66	3.65	9.07	2.82
	Avg.	1.42	28.10	8.06	122.42	41.89	12.21	3.34	7.18	2.95

Figure 4.3 *HOX* is mildly over-expression in the chMM.
A) Schematic representation of an RCASBP(A) vector in a linear form. Three possible splice variants are indicated and the position of the viral and *HOX* genes is accurately portrayed to comprehend the splice variants. B) A schematic depiction of the posterior, distal part of the HH25 forelimb in the chick embryo that was dissected for the comparison of the transcript abundance between the *HOX* expressing tissue and chMM over-expressions. C) Quantification of the over-expression level in the chMM in comparison to the *in vivo* tissue of expression. Top panel: Splice variant quantification from the chMM RNA-seq. Approximately one third of total viral RNA leads to the transcript that carries *HOX* and will have a viable protein product. Middle panel: RPKM values for all HOX genes in their own overexpression and for the tissue control experiments, in replicates. Bottom panel: Quantification of the over-expression level for the each of the two replicates and below the average over-expression level. D) Western Blot of the over-expression chMM cultures for each of the nine HOX. Loaded protein was extracted from the same number of cells and quantified by Bradford assay to ensure comparability.

First, chMM RNA-seq was filtered only for the viral mRNA variant that leads to the production of the HOX protein. This accounted for 15%-45% of the total viral mRNA (**Figure 4.3C**). From this, the chMM *HOX* gene RPKM values were calculated.

At the same time, *in vivo* HOX gene RPKM from the posterior, distal forelimb were calculated. From these numbers, an overexpression level was easily inferred and ranged from 1,42 fold for the *HOXA9* gene up to 122 fold for *HOXA13* gene. Having in mind that *HOXA13* gene was very poorly expressed in the microdissected tissue it was expected that it would show the highest overexpression. However, when taking into account that only approximately 20% of the cells expressed *HOXA13*; it is clear that the overexpression level is reduced significantly. Importantly, *HOXD11* gene, which was best represented in the microdissected tissue showed merely three

fold overexpression. These results corroborated overexpression folds calculated in other studies using the same system (Ibrahim, 2014).

Lastly, to make sure that transcriptional differences between overexpressing cells and *in vivo* embryonic tissues are a good proxy for the protein overexpression in the chMM, a Western Blot analysis was performed. Here, the same amount of total protein was loaded for every chMM overexpression and visualized on the blot using the same antibody. Importantly, chMM displayed a uniform expression level amongst all nine HOX proteins.

Altogether, the overexpression fold induced by RCASBP(A)-*HOX* constructs are quantifiably mild, both on transcript and protein level. Therefore, the chMM system with this kind of overexpression was deemed suitable to use for subsequent HOX-TFs binding and functional analyses.

4.4 Characterization of Regulatory Programs Induced by Individual HOX-TF Overexpression

4.4.1 Comparative Analysis of Induced Regulatory Programs

Overexpression of *HOX* genes in the chMM system offers a lot of possibilities to study the behavior of TFs and its' binding, but also, to some degree, to study induced transcriptional effects. Although a cell culture approach is difficult to compare to an *in vivo* situation, this system offered a unique opportunity to study transcriptional programs induced by an individual *HOX* gene which is impossible *in vivo* due to their overlapping expression patterns and functional redundancy.

General Transcriptional Analysis

First, RNA was collected from the chMM cultures at day six post-infection. This time point was selected to ensure high cellular infection rate. RNA was then sequenced, mapped and differentially regulated genes were generated. To do so, RNA-seq from every HOX overexpression culture was compared with RNA-seq from an uninfected control chMM culture. Differentially regulated genes were filtered according to the p-value, base mean expression and fold change. Total of 166 and 213 genes were found to be differentially regulated for the overexpression of the *HOXA10* and *HOXA11* genes, respectively (**Figure 4.4A**). However, *HOXA13* and *HOXA9* overexpressions generated a much higher number of differentially

regulated genes than the rest of the *HOX* overexpressions. *HOXA13* gene overexpression regulated a total of 601 genes and *HOXA9* 598 genes.

Next, I compared induced transcriptional programs. For this, top 50 differentially regulated genes from every *HOX* induced culture were selected and clustered (**Figure 4.4A**). Clustering detected several significant findings. First, *HOX* genes induce similar transcriptional programs and share many target genes. Here, induced transcriptional programs are quite redundant since only 206 genes were found in the top 50 differentially regulated genes among nine datasets. This result demonstrated that functional redundancy is present at the individual targets as well as on a gross anatomical level. Additionally, hierarchical clustering uncovered some associated features reminiscent of *HOX in vivo* functions. Exemplarily, HOXD13, and HOXD12 transcriptional programs cluster closer than the PG10, PG11, and HOXD9 which is reminiscent of their roles in the limb patterning, where HOXD13 and HOXD12 TFs act in autopod and HOXD9, PG10, and PG11 in zeugo-/stylopod,(**Figure 4.4A**). Second, although, *HOXA13* and *HOXA9* induced more distinct regulatory programs, their strongest regulated target genes were common target genes regulated by other *HOX* genes as well (**Figure 4.4A**).

Finally, a closer examination of target genes uncovered these genes were primarily proliferation- and differentiation-related genes, with almost no known developmental TFs present. Then, differentially regulated genes were subjected to the gene ontology (GO) analysis to identify major processes induced by overexpressions. Expectedly, the GO analysis mainly identified terms related to skeletal development, vertebrae transformation, and anatomical processes, as these processes are known to depend on *HOX* genes (**Figure 4.4B**).

HOX Autoregulation

Due to the complex *HOX* gene expression pattern little direct evidence is available to describe HOX autoregulation. Therefore, I decided to examine the chMM transcriptomic data to identify if any HOX autoregulation is detectable. For this purpose, RPKM values were calculated for each of nine *HOX* genes in the overexpression experiments, in comparison to the control experiment (**Figure 4.4C** and **D**). This analysis uncovered an interesting tendency for posterior gene upregulation. More specifically, the overexpression of every tested *HOX* gene, except *HOXA9*, leads to *HOXD13* and *HOXA13* upregulation (**Figure 4.4C** and **D**). These findings suggest a trend where protein products of more anterior HOX (e.g. HOXD10) genes positively affect the expression of more posterior genes (e.g. *HOXA13* and *HOXD13*).

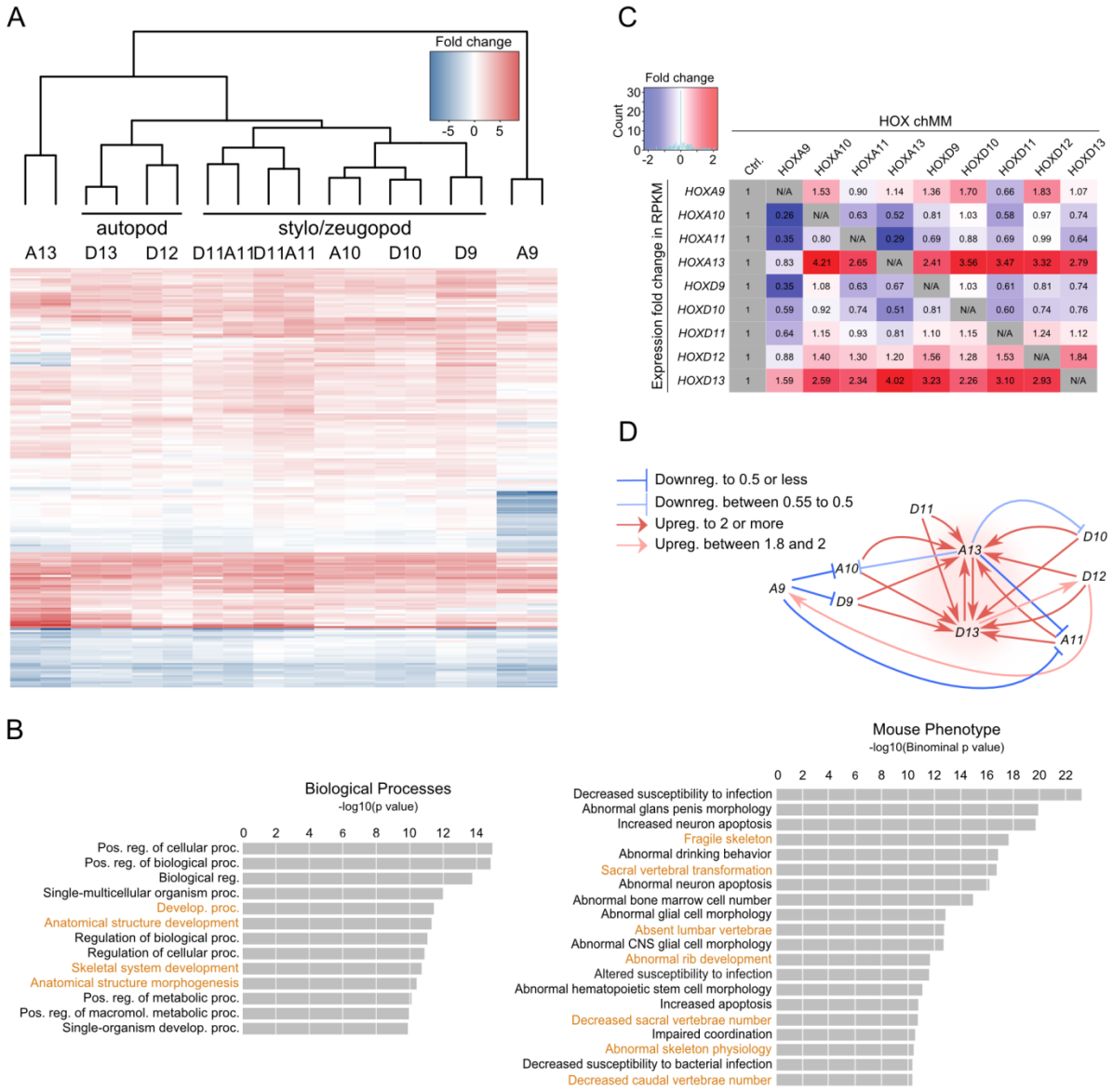


Figure 4.4 HOX induce similar transcriptional programs.
A) Hierarchical clustering and dendrogram for every gene in replicates. Following parameters were used: top 50 deferentially regulated genes for every sample, base mean≥30, log2fold change≥1, pvalue≤10^{-5}. B) The RPKM values were calculated for the *HOX* genes in all the over-expression experiments. RPKM values are colored according to regulation. Most down-regulated showing in deep blue to the most up-regulated showing in bright red. C) A schematic representation of the HOX autoregulation based on the data in B). Color of the arrows represents either the up-regulation in red or down-regulation in blue. D) Gene Ontology (GO) analysis with GOrilla and GREAT tools from left to right, respectively (Eden *et al.*, 2007, 2009; McLean *et al.*, 2010). The input genes for these tools were same in A).

Together, these findings showed that a single *HOX* gene overexpressions are sufficient to induce redundant transcriptional programs. Furthermore, no single posterior *HOX* gene overexpression can induce the change in expression of well-known developmental master regulators. In this view, the impact of individual *HOX* gene expression is relatively limited and is likely affecting

biological processes more extensively when expressed in combination with other paralogues. Lastly, *HOX* genes displayed a degree of autoregulation within and between the clusters, suggesting an even more complex situation when several *HOX* genes are expressed in the same tissue/time.

4.4.2 Paralogy Group (PG) Specific Regulatory Programs

HOX genes from the same paralogy group are remarkably homologous and functionally redundant (Zakany & Duboule 2007). Therefore, next, I investigated functional redundancy within individual paralogy groups. For this, every HOX induced transcriptional programs were directly compared to its paralogue's (e.g. HOXA10 vs. HOXD10). This comparison uncovered that PG10 and PG11 target almost all of the same genes, with only 7 and 11 differentially regulated genes within their groups, respectively (**Figure 4.5B**). Conversely, the PG9 and PG13 displayed most variability in their respective paralogy groups with 567 and 263 differentially regulated genes, respectively (**Figure 4.5B**). These results indicate that some paralogy groups are more redundant than others, at least in this system. Interestingly, a comparison between HOXD13 and HOXD12 induced transcriptional programs identified only 30 differentially regulated genes, corroborating previously defined functional redundancy between these genes (Kmita et al. 2002; Kmita et al. 2005).

Finally, GO analysis of the PG9 and PG13 differentially regulated genes was performed to determine the processes associated with these genes. While both, PG9 and PG13, impact similar features like the anatomy of the limb, vertebrae, and autopod; the cellular components which are affected by these processes differ. PG9 group of proteins is enriched for the terms like PRC1 and PcG complex whereas the PG13 group affects processes mainly associated with extracellular matrix (**Figure 4.5C**). This further describes the differences within these paralogy groups and indicates that even apparent discrepancy in the targeted genes somehow could lead to the similar phenotypic consequence.

Taken together, the paralogue-specific transcriptomic analysis showed that while there is a general redundancy within the *HOX* induced transcriptional programs, the extent of the redundancy is individual to paralogy groups. In such way, PG10, PG11, and to some extent the neighboring *HOXD13-HOXD12* induce many same targets, whereas PG9 and PG13 differ extensively. GO analyses between the more diverse paralogy groups displayed puzzling plasticity

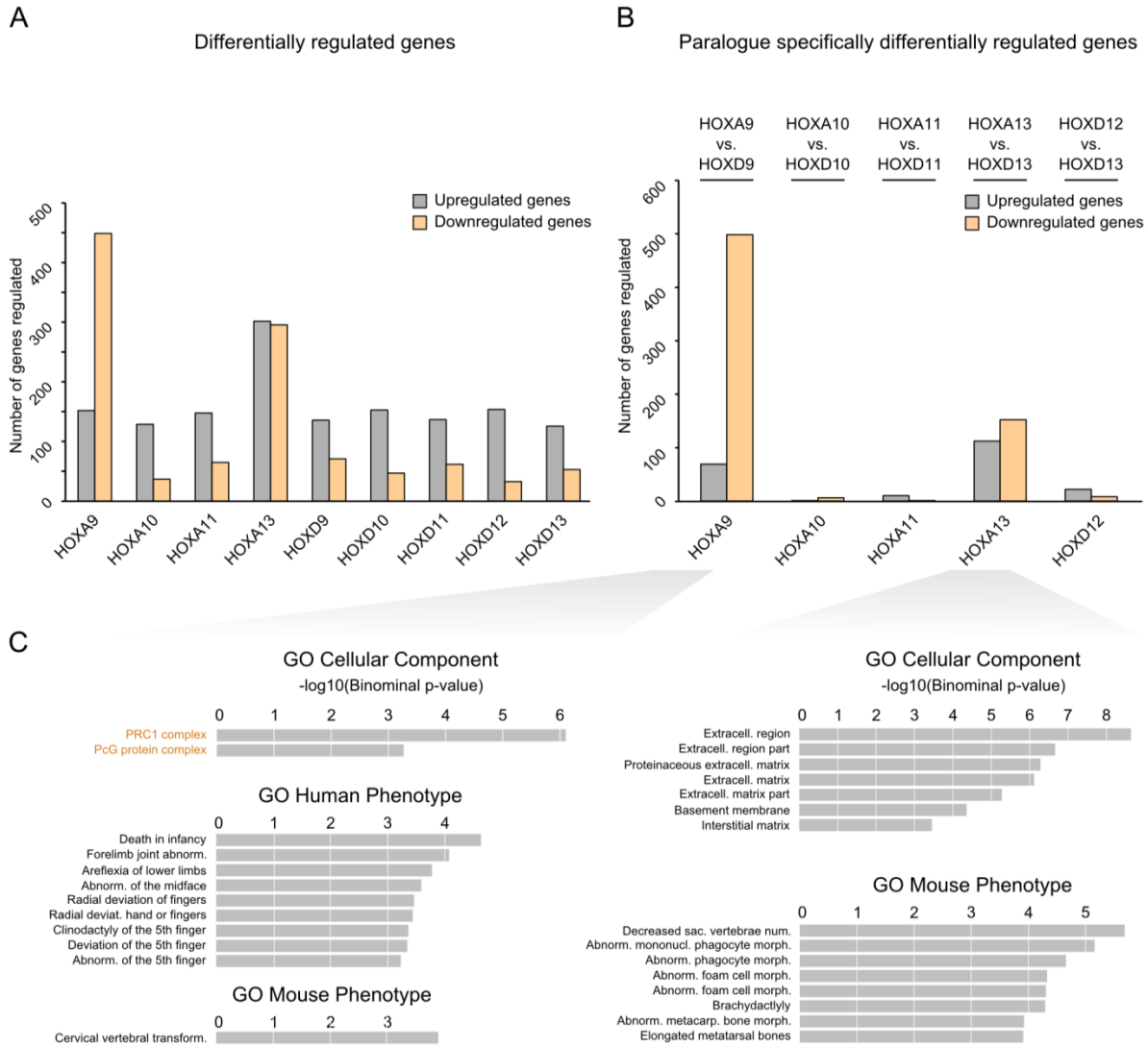


Figure 4.5 PG10 and PG11 induce nearly identical regulatory programs.
A) All up-regulated and down-regulated genes quantified for every sample. Deferentially regulated genes were selected using the following parameters: base mean ≥ 100, log2fold change ≥ 1, p-value ≤ 10-5. B) Paralogy group specific comparison of deferentially regulated genes. Parameters used as in B). C) GO analysis of the deferentially regulated genes from the PG9 and PG13 using GREAT (McLean *et al.*, 2010).

as quite distinct transcriptional programs seem to be linked with similar associated processess and phenotypic outcomes.

4.5 Analysis of HOXA/D TF Binding

DNA binding is most essential HOX property. However, it is largely understudied due to the technical limitations. Using chromatin immunoprecipitation followed by the next generation sequencing (ChIP-seq), I generated HOX genome-wide binding profiles with few hundred base pair resolution. This technique allows not only the localization of the protein binding but also for better understanding of the binding mechanism and discovery of putative cofactors. Here, I employed ChIP-seq to map, investigate, and compare genomic binding profiles of HOX-TFs.

4.5.1 Initial Analysis and Validation of HOX-TF ChIP-seq

For a successful ChIP-seq, it is imperative to have a highly specific antibody. For HOX, this was for a long time the most limiting factor. Here, however, this problem was circumvented with a design of a fusion gene that contains an N-terminal, unique epitope followed by the *HOX* gene of interest. Still, even with all the requirements met, performing a high-quality ChIP-seq is not trivial and one must thoroughly examine ChIP quality and reproducibility. For this, I used an established pipeline that was designed and implemented according to the ENCODE guidelines by Dr. Daniel M. Ibrahim and Mr. Peter Hansen (Li *et al.*, 2011; Landt *et al.*, 2012).

Two most useful criteria for measuring the quality of ChIP-seq are high enrichment and reproduciblity between biological replicates. First, ChIP-seq enrichment quality is measured for every ChIP-seq replicate and visualized with the SPP plot (**Figure 4.6A**) (for details see **Chapter 3.6.3**). All here performed ChIP-seq results passed the initial analysis. However, few samples (e.g. HOX9 and HOX10) scored somewhat lower on the enrichment analysis even after performing three biological replicates. For these samples, I followed ENCODE guidelines that require slightly marginal ChIP-seq to be checked for reproducibility and found they are highly reproducible despite marginal enrichment (**Appendix 1**). Then, as suggested by ENCODE guidelines, for these TFs I chose two best performing replicates and continued the downstream analysis (Li *et al.*, 2011; Landt *et al.*, 2012).

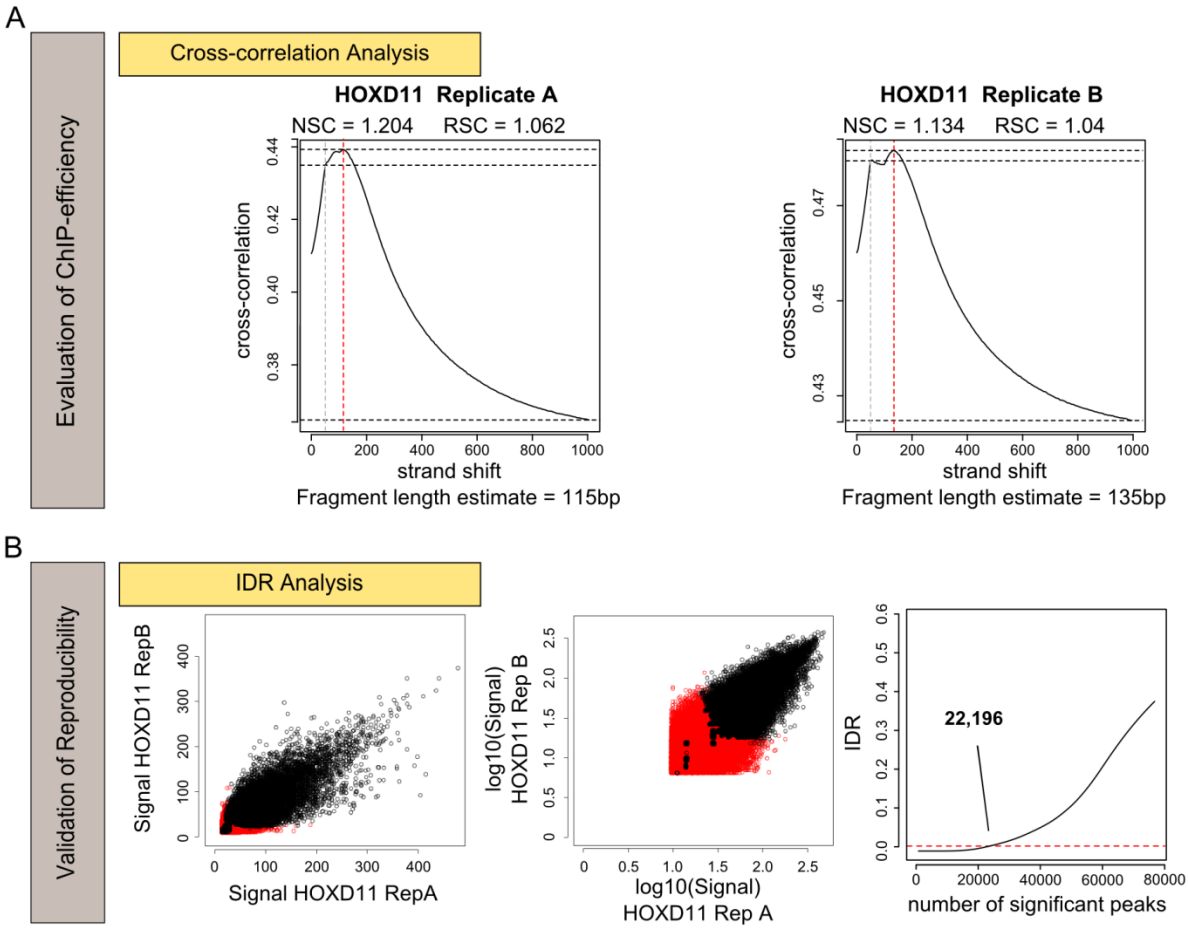


Figure 4.6 HOXD11 ChIP-seq dataset are reproducible.
A) SPP cross-correlation analysis of the two HOXD11 replicates, Replicate A and Replicate B indicating the signal to noise ratio (NSC and RSC), fragment length estimate and strand shift. B) Irreproducibility Discovery Rate analysis (IDR) checking reproducibility of data between two biological replicates (Black dots - reproducible peaks, red dots - irreproducible peaks). The diagram on the far right depicts the number of called peaks above the selected threshold. All the peaks discovered under the threshold (dotted red line) are reproducible peaks (Li *et al.*, 2011; Landt *et al.*, 2012).

Second, ChIP-seq samples were performed in biological replicates which had to meet the reproducibility requirements (irreproducibility discovery rate, IDR)(for details see **Chapter 3.6.3)**. Having two highly reproducible experiments is crucial since ChIP-seq is an experiment with many wet-lab variables and very few checkpoints. Here, the IDR analysis for every ChIP-seq experiment verified the biological reproducibility of binding sites (**Figure 4.6B** and **Appendix 1**). Furthermore, IDR analysis also allowed for the discovery of reproducible binding sites and suggested the number of high-confidence, reproducible peaks (**Figure 4.6B** right panel) (for details see **Chapter 3.6.3**). With this pipeline, the number of reproducible peaks for HOX-TFs ranged between 13 236 and 29 261 for HOXA13 and HOXA10, respectively (**Appendix 1**).

Furthermore, peaks mapping to mitochondrial DNA or the poorly assembled parts of the chicken genome were excluded from any subsequent downstream analysis of the ChIP-seq data.

4.5.2 Identification of Posterior HOXA/D TFs Binding Sites

Posterior HOXA/D are structurally very similar as demonstrated in **Chapter 4.1**. Despite this, HOX-TFs, particularly paralogy groups PG9 and PG13, demonstrated quite unique induced regulatory programs as described in **Chapter 4.4.** Findings presented in these two chapters demonstrate internally detected HOX paradox. To gain insights into HOX paradox in the following chapters I analyzed and compared HOX-TF binding profiles, their canonical, and noncanonical binding sites.

First, HOX ChIP-seq data were mapped and displayed using UCSC browser to visualize binding and to aid the later interpretation of results (**Figure 4.7A**). A quick browsing of data indicated that much like proteins themselves, binding patterns are partially redundant (**Figure 4.7A**). However, while many genomic locations harbored binding of several HOX-TFs, binding of all nine HOX-TFs at the same location was a rarer event. Additionally, some genomic locations were bound only by one HOX-TF or only one paralogy group. To compare these binding profiles in an unbiased way, I collected genome-wide binding sites and performed Principal Component Analysis (PCA). PCA resulted in a major principal component (PC1), describing 37% of the variance in the data, and several smaller components, most prominently principal component 2 (PC2) representing some 13% variation in the data. Two major principal components were then plotted to compare HOX binding profiles (**Figure 4.7B**). PCA analysis uncovered two major regiments. First, PG13 binding pattern are separated from other HOX-TF binding along the PC1, indicating there is a subset of binding sites only bound by PG13. Second, an unusual and novel subgrouping appeared along PC2 which did not correlate with any known HOX property or function (**Figure 4.7B**). This novel subgrouping separated HOX-TF binding in two groups, which were named: Group 1 (HOXA11, HOXD11, and -D13); and Group 2 (HOXA9, -A10, -A13, HOXD9, -D10, and -D12).

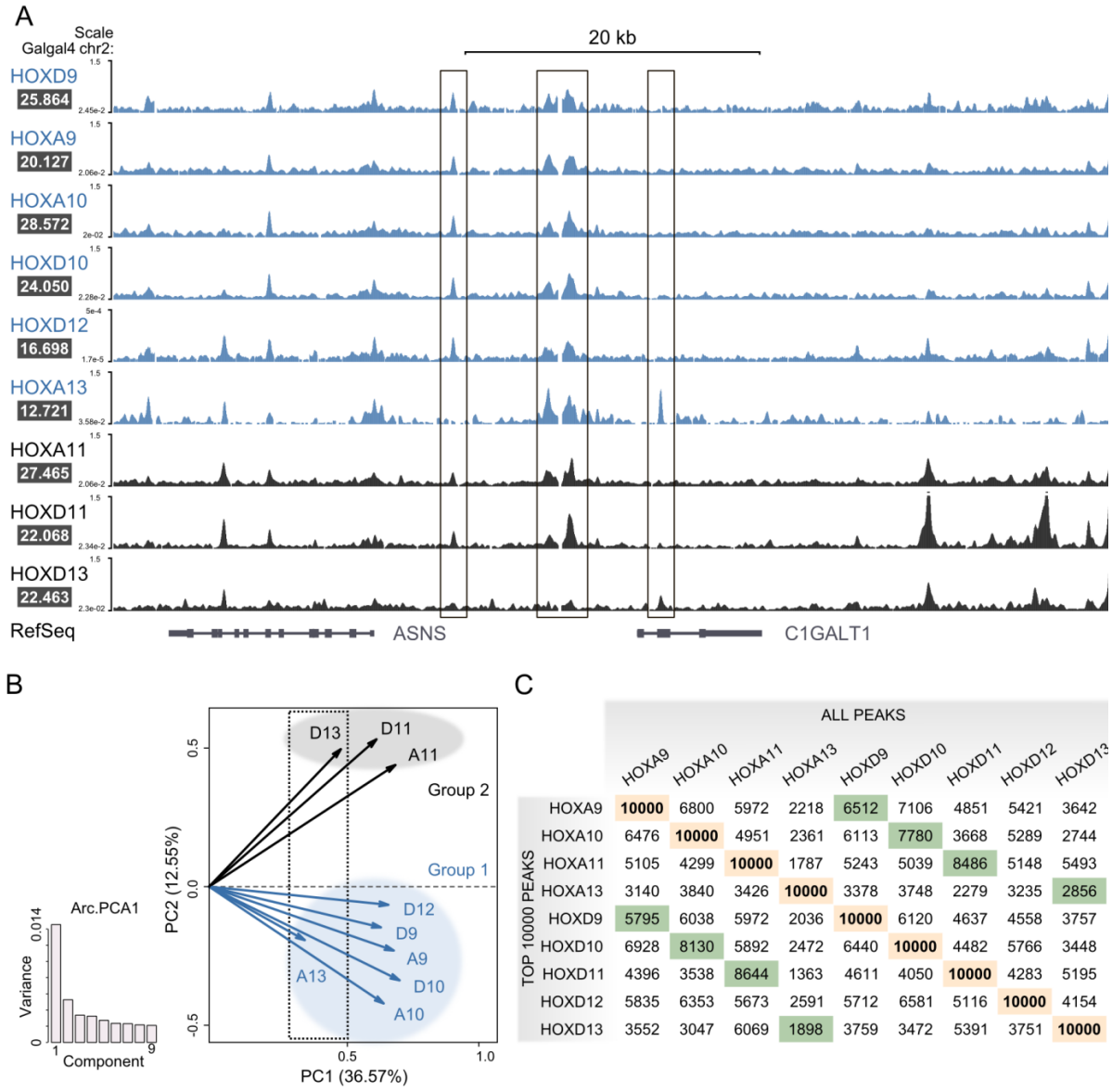


Figure 4.7 Posterior HOXA/D uncover novel sub-grouping according to their genome-wide binding profiles.
A) UCSC tracks of the HOX ChIP-seq profiles at a random region in the chicken genome. B) Principal Component Analysis (PCA) of the first two components based on the HOX binding. C) Pairwise analysis of the HOX co-occupancy of same genomic regions.

Next, I performed a pairwise analysis between each HOX-TF, primarily to investigate the binding redundancy within the paralogy groups. For this, top 10 000 peaks of every dataset were crossed with all peaks of the second dataset since a total number of identified peaks varied between nine datasets (**Figure 4.7C**). This analysis revealed that paralogy groups are often highly redundant sharing, 19-29% binding sites within PG13 (HOXA13-HOXD13), 58-65% within PG9

(HOXA9-HOXD9), 78-81% within PG10 (HOXA10-HOXD10), and 86-85% within PG11 (HOXA11-HOXD11) (**Figure 4.7C**). The observed reduction of the PG13 common binding can be partially attributed to a poorer reproducibility of the HOXA13 ChIP-seq replicates. However, a correlation between binding redundancy and transcriptional redundancy indicates that HOX biological properties are a more likely culprit. Specifically, PG9 and PG13 show less binding redundancy and less redundancy in their induced regulatory programs. In contrast, the PG10 and PG11 show highly redundant binding and highly redundant induced regulatory programs. While these observations are purely associative, they indicate that observed binding sites represent true binding events with functional consequences.

This study generated the first record of vertebrate, *in vivo*, genome-wide binding maps of posterior HOXA and HOXD-TFs demonstrating that posterior HOX-TFs sometimes tend to bind same genomic locations. This is especially prominent within some paralogy groups (PG10 and PG11), but less so within other paralogy groups (PG9 and PG13). Finally, binding profiles revealed a novel and unexpected subgrouping that could not be justified by any known functional or biochemical property.

4.5.3 Functional Validation of Posterior HOXA/D TF Binding

Transcription factors, like HOX, most often bind at distal regulatory elements. Therefore, it is a challenging task to functionally validate their binding. For this purpose, it is necessary to investigate functionality of individual binding events in detail. However, such approach is time consuming and would lead the analysis to a target specific approach. A genome-wide alternative approach is to use existing functional database and investigate binding at known limb-specific enhancers.

For this purpose, I used the VISTA enhancer browser (Visel *et al.*, 2007). VISTA enhancer browser is an online resource of experimentally validated human and mouse *cis*-regulatory elements. First, human and mouse limb-positive and fore/midbrain-positive enhancers were selected. Limb-positive enhancers were used as a positive functional validation of HOX-TF binding and fore/midbrain positive enhancers were used as a negative control. Furthermore, fore/midbrain positive enhancers were additionally screened to exclude enhancers active in any *HOX*-expressing tissues, to prevent false positives. After successful enhancer selection, coordinates were lifted over from human or mouse onto the chicken genome. Since VISTA Enhancer Browser selects *cis*-regulatory elements for functional testing by conservation, it was

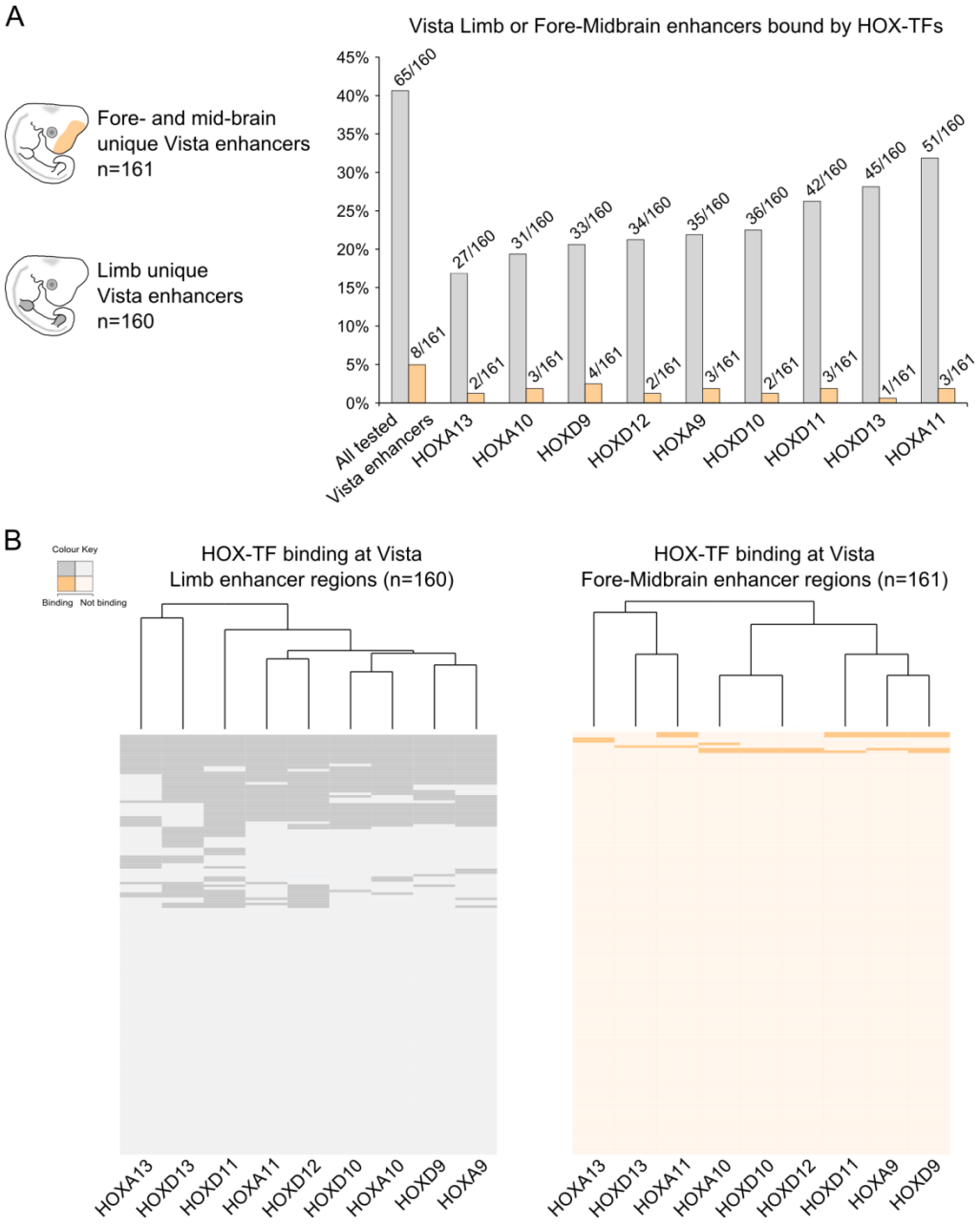


Figure 4.8 HOX bind to 41% of all mouse and human VISTA-tested limb enhancers.
A) HOX binding to the limb-positive enhancers, and fore- and midbrain positive enhancers. The enhancers were lifted over from either the mouse or human genome to the chicken Galgal4 genome and then the binding was tested. B) Clustering of the HOX binding at limb or brain enhancers.

possible to lift over a majority of VISTA enhancers. A total of 160 limb-positive enhancers and 161 of fore/midbrain-positive enhancers were successfully lifted over. Then, HOX-TF binding at these enhancers was tested. A total of 41% (65/160) of limb-positive enhancers was bound by one or more HOX-TFs, whereas only 5% (8/161) of fore/midbrain enhancers were bound (**Figure 4.8A**). More specifically, individual HOX binding was present at 27 and 51 limb-specific

enhancers for HOXA13 and HOXA11, respectively. For comparison, HOX-TF binding at the fore/midbrain enhancers ranged between one and up to the four enhancers, in total.

To investigate this in more detail, HOX binding at these enhancers was clustered (**Figure 4.8B**). Interestingly, about half of HOX-positive limb enhancers were bound by at least two different HOX-TFs. Therefore, there are two subsets HOX-bound limb enhancers, the commonly occupied enhancers, and more specific enhancers. Finally, the dendrogram representing the HOX-bound limb enhancers indicated the paralogy groups often occupy same enhancers (**Figure 4.8B**). Specifically, even though HOX-TFs commonly bind many of limb-positive enhancers, PG9, PG10, and PG13 bind most of the same enhancers (**Figure 4.8B**).

These results functionally validate HOX-TF binding sites identified in this study. Furthermore, it also indicates there are distinct subsets of HOX-binding limb enhancers, common and specific, leaving room for speculation on the possible roles of these two tiers of HOX-bound enhancers.

4.5.4 Primary Motif Analysis

Transcription factor binding is often best described by motifs overrepresented in ChIP-seq data. Binding sites can contain a diverse set of motifs, primary and secondary, which describe different binding modes of the investigated TF. Exemplarily, overwhelming and predominant presence of a motif (primary motif) belonging to the ChIP'ed protein means that most of the sites likely represent direct binding. Conversely, presence of slightly changed primary motif can infer the cofactors' influence on the TF-DNA binding. HOX-TFs are particularly notorious for their dependency on the cofactors. So far, this phenomenon has been studied, with large *in vitro* screens on protein binding array (PBM) and SELEX (Berger et al. 2008; Jolma et al. 2013; Slattery et al. 2011; Jolma et al. 2015). Alternatively, *in vivo* cooperative HOX binding has been studied only on several loci in *Drosophila* (Crocker *et al.*, 2015). However, no data exists to describe *in vivo*, genome-wide binding of HOX-TFs.

Here adapted chMM system, allowed investigations of genome-wide HOX-DNA binding in the presence of the cofactors, and in an adequate cellular environment. Therefore, the first aim was to identify *de novo* motifs present in HOX ChIP-seq datasets and to infer binding modes from these data.

To investigate the overrepresented binding motifs top 5 000 peak summits (151bp length each) were used as an input for *de novo* motif analysis (RSAT tool) (**Figure 4.9A**). Most strongly bound

peaks were selected since these peaks often represent direct binding. Thus, primary motifs describing direct binding will be most abundant. Top three overrepresented motifs were identified and reported, except for HOXD9 where only two overrepresented motifs were found (**Figure 4.9A**). Furthermore, HOXD9 and HOXA9 each uncovered one motif that was known to be a composite site of HOX and a cofactor. Interestingly, an HOXA9 non-HOX motif, TTCATGAA was closely matched to an HOXA9-PBX1 motif identified by crystallography (**Figure 4.9B**). This finding illustrates the importance of latent specificity for the HOX-DNA binding and demonstrates it is possible to identify specific *in vivo* motif changes using this system and analysis (LaRonde-LeBlanc 2003; Slattery et al. 2011).

Up-to-date any and all available HOX-DNA binding data derives from *in vitro* analyses. Therefore, oligo motifs that bind to Homeodomain with the highest affinity represent data that derived from these *in vitro* analyses (**Figure 4.9A**, right panel). In comparison to these published data, Group 1 HOX had greater tendency to change primary motifs, especially the typical HOX, –TAAA 3' part of the motif. The entire Group 1 had this –TAAA 3' part changed to a –TGAA 3'. Furthermore, 5' part of the sequence was more prominent and changed from the N[C/T][C/A]A to G[C/T]AA (**Figure 4.9A**).

Group 2 also had changed motifs but to varying degree. HOX11-TFs exhibited changes in the entire motifs while HOXD13 motif was minimally changed. HOX11 paralogues had a more versatile 5' part of the motif which was changed from an *in vitro* G[T/C][C/A][G/A to an *in vivo* G[T/C][A/C]A (**Figure 4.9A**). Group 2 motifs also had well conserved 3' –TAAA tail, although this tail was sometimes interchanged with the –TGAA (**Figure 4.9A**). Lastly and importantly, binding of the PG13 proteins demonstrated these two proteins are most commonly bound by their own canonical, monomer-like motifs (**Figure 4.9A**). Furthermore, HOXA13 and HOXD13 have been reported, to indistinguishably bind both their own and each other's canonical motifs (Taneda *et al.*, 2004; Zhang *et al.*, 2011; Jolma *et al.*, 2013, 2015; Turner *et al.*, 2014).

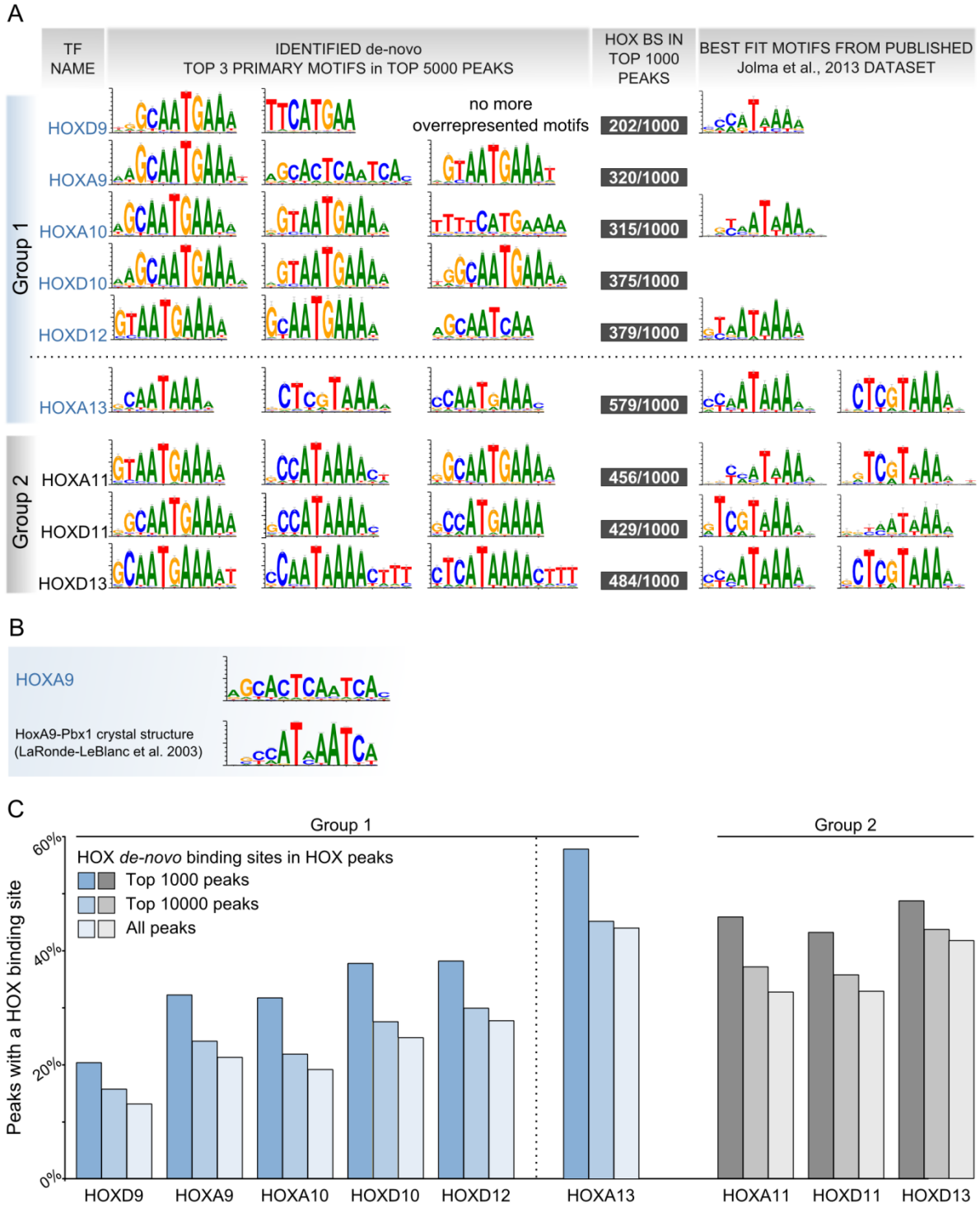


Figure 4.9 *De-novo* motif analysis uncovers distinct changes in binding preferences.
A) *de novo* motif analysis of top 5 000 peaks using RSAT tool (Thomas-Chollier *et al.*, 2012). B) Comparison of the HOXA9 second motif to the PBX1-HOXA9 composite site motif using TOMTOM tool (Gupta *et al.*, 2007). C) Quantification of the HOX binding through its primary motifs (direct binding) using FIMO tool (Grant, Bailey and Noble, 2011).

Similarly, in this study, *de novo* motif analysis confirmed that both HOXD13 and HOXA13 bind their own and each other's canonical motif CCAATAAA and CTC[G/A]TAAA, respectively (**Figure 4.9A**). Finally, I found the canonical 3' -TAAA tail of the PG13 motifs to be more conserved within this group than for the rest of the HOX-TFs.

Primary motifs demonstrated cofactors have a substantial impact of on the direct HOX-DNA direct binding, except for the HOX13 TFs. To understand how many sites rely on this binding mode, it was necessary to quantify this direct binding. Using FIMO, all of three primary *de novo* discovered motifs were quantified in their respective HOX-TF binding sites. Quantification of the direct binding demonstrated three main things. First, the **direct HOX-DNA binding is low**. Only 20%-58% of the top 1 000 peaks contained primary motif for HOXD9 and HOXA13, respectively (**Figure 4.9C**). This number decreases in the top 10 000 peaks to 16%-45% for HOXD9 and HOXA13, respectively (**Figure 4.9C**). In all peaks, direct binding is reduced down to 13%-44% for HOXD9 and HOXA13, respectively (**Figure 4.9C**). It is expected to see a decline in the number of the primary motif as the search expands to all peaks since the most strongly bound peaks are most often are bound directly. Second, **Group 1 demonstrated less direct binding than Group 2**. This is accompanied with Group 2 binding more often through canonical sites than Group 1, which is particularly the case for HOX13 paralogues. Third, analysis of direct binding suggested that **most of HOX-DNA binding does happens indirectly** through tethering or indirectly through some other mechanism.

Altogether, *de novo* motif analysis uncovered presence of altered, noncanonical motifs for all HOX-TFs. Alterations in primary motifs partially explain subgrouping discovered in the PCA analysis, but cannot explain HOXA13 presence in the Group 1, nor what characterizes Group 2. Direct binding quantification further corroborated these findings and suggested direct binding as a partial driver of Group 1 and Group 2 discrepancy.

4.5.5 Secondary Motif Identification

De novo motif analysis is a useful tool to uncover which motifs are overrepresented in binding data. The advantage of such analysis is its unbiased nature as such approach will discover any over-represented motif. Since previous analyses indicated that HOX-TFs do not always bind directly, it was necessary to investigate which possible other TFs might be involved. First, all peak summits were (151bp each) subjected to RSAT motif analysis (Thomas-Chollier *et al.*, 2012). This search identified the presence of several non-HOX motifs in Group 1, but not Group 2, which is

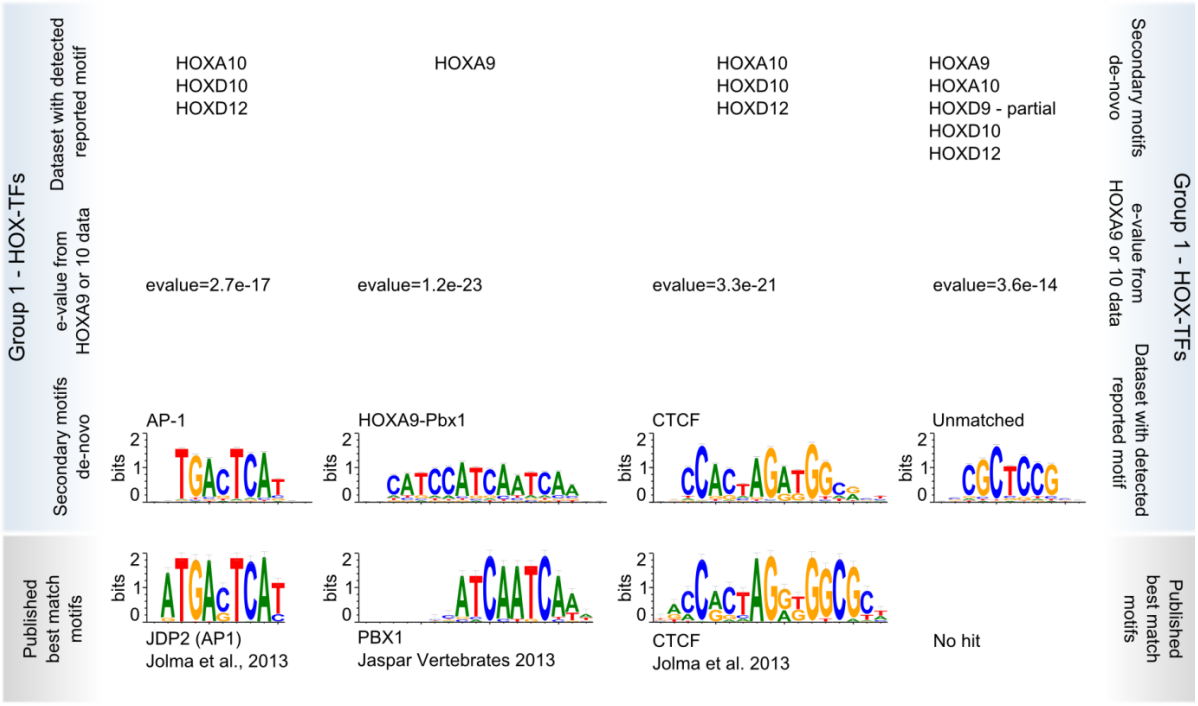


Figure 4.10 Expanded *de novo* motif analysis uncovers three major classes of secondary motifs. TOP: HOX data where *de novo* motif was discovered and e-values. MIDDLE: *de novo* discovered motifs. BOTTOM: best match motifs from database.

partly due to the fact that Group 1 HOX-TFs do not often bind DNA directly. Furthermore, no secondary motifs were discovered in HOXA13 binding data indicating that strong presence of primary motifs might be overshadowing *de novo* motif search for other motifs. In brief, this analysis uncovered four different secondary motifs: an AP1 motif; an HOXA9-PBX1 motif that was also discovered in the previous motif analysis; a CG-rich motif that was not matched with any TF; and a CTCF motif (**Figure 4.9B**, **Figure 4.10** and **Appendix 2**) (Gupta *et al.*, 2007; Thomas-Chollier *et al.*, 2012; Mathelier *et al.*, 2014).

4.5.6 Quantification and Verification of AP1 and "Unmatched" as secondary Motifs

HOX binding that is impoverished of direct binding is necessary to investigate as it is informative of other than direct TF-DNA binding. Above presented descriptive analysis of secondary motifs was a first step to identify potential cofactors or tethering factors. However, it is necessary to investigate these motifs in more detail as they can be present either at a background level

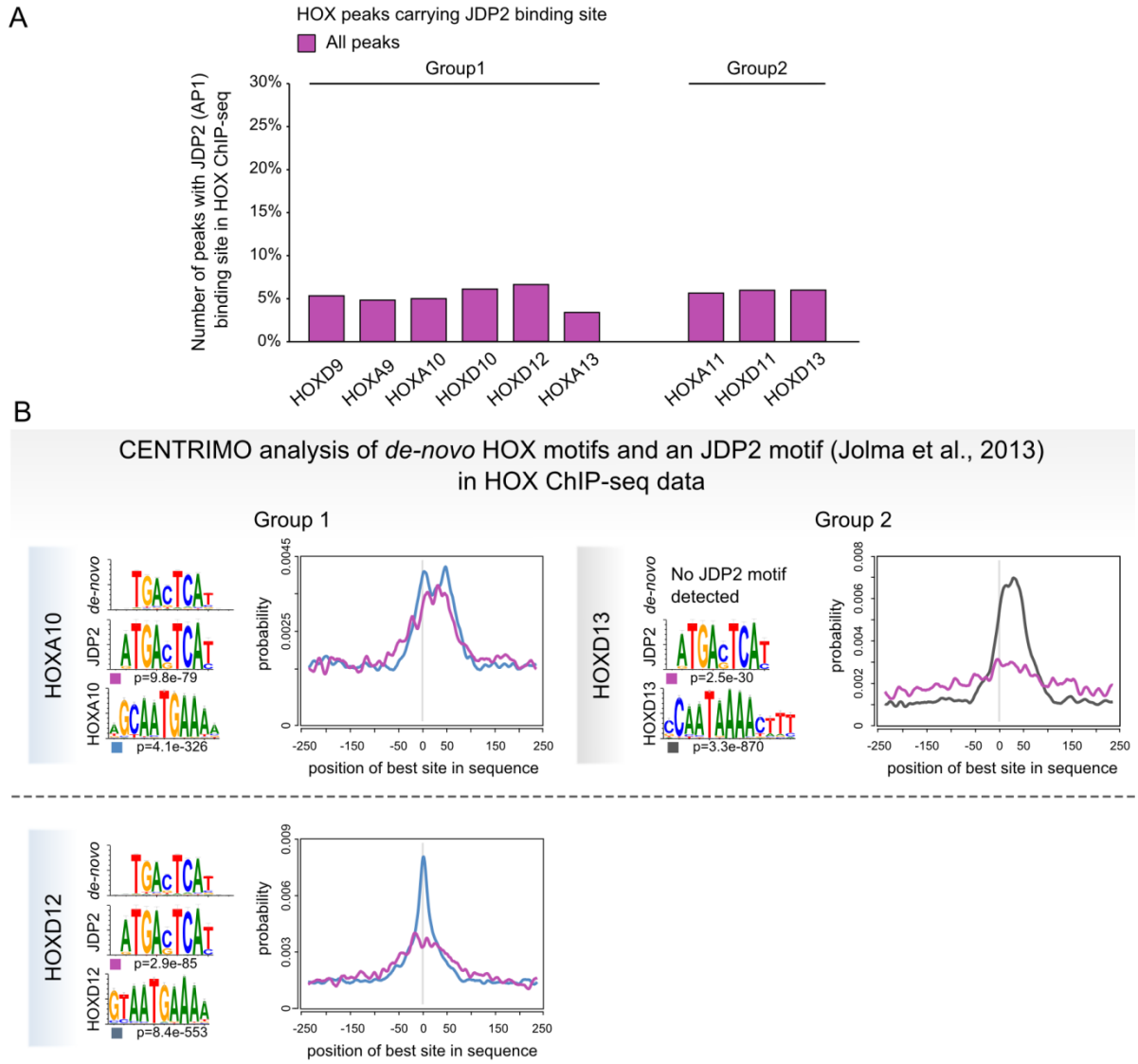


Figure 4.11 AP1 motif is present in HOX binding data as a secondary motif but at a very low amount.
A) Quantification of the AP1 motif in HOX data with FIMO tool (Grant, Bailey and Noble, 2011). B) Probability of finding the AP1 motif and AP1 motif positioning in HOX binding data. Analyzed with Centrimo tool (Bailey and MacHanick, 2012).

throughout the binding site or they can be enriched at the peak summit like the primary motif would be.

Since true secondary motifs, unlinked to HOX[4], are only AP1, "Unmatched", and CTCF subsequent analysis was focused on these motifs. In order to investigate motifs enriched in each of the HOX peaks two software were used, Centrimo and FIMO (Grant, Bailey and Noble, 2011; Bailey and MacHanick, 2012). Centrimo detects localization of the motif in peak and the probability of finding the motif in respect to peak summit. On the other hand, FIMO allows for the detection and quantification of specific motifs.

AP1 Motif Investigation

To examine AP1 motif, an AP1 motif position weight matrix (PWM) was extracted as identified in the HOXA10 dataset. Then, all nine HOX-TF binding sites were investigated for the central enrichment of the AP1 motif. Centrimo analysis demonstrated AP1 motif was present at HOX peaks with varying probability in comparison to the primary motif (**Figure 4.11** and **Appendix 3**). Only the PG9 and PG10 had an AP1 motif present around the peak summit (**Appendix 3**). Furthermore, AP1 motif was present in all HOX data in about only 3-6% of the peaks (**Figure 4.11A**). These analyses demonstrated that this TF motif is often present at varying positions and is not highly enriched in HOX binding data.

"Unmatched" Motif Investigation

Next, the "Unmatched" motif was subjected to the same analysis as the AP1 motif. First, the "Unmatched" motif was found with low probability around the summit of all HOX-TF binding, with the notable exception of the HOXA13 where this motif was found to be centrally positioned at the summit site and with high probability (**Figure 4.12B** and **Appendix 4**). This motif was slightly more abundant in Group 1 than in Group 2 (**Figure 4.12A**). However, the importance of the "Unmatched" motif was disputed with Centrimo analysis since it was not detected centrally at the peak summit.

Lastly and importantly, none of the above holds true for the HOXA13 which is the only HOX-TF where the "Unmatched" motif is present at the peak summit and quite abundantly as well (**Figure 4.12B** and **Appendix 4**). It is unclear whether this is a motif of another TFs or the HOXA13 protein is biased to these sites by some other mechanism and this motif is a secondary

[4] HOXA9-PBX1 motif is not investigated here in more detail as it is known to be a feature of TALE driven, altered direct binding.

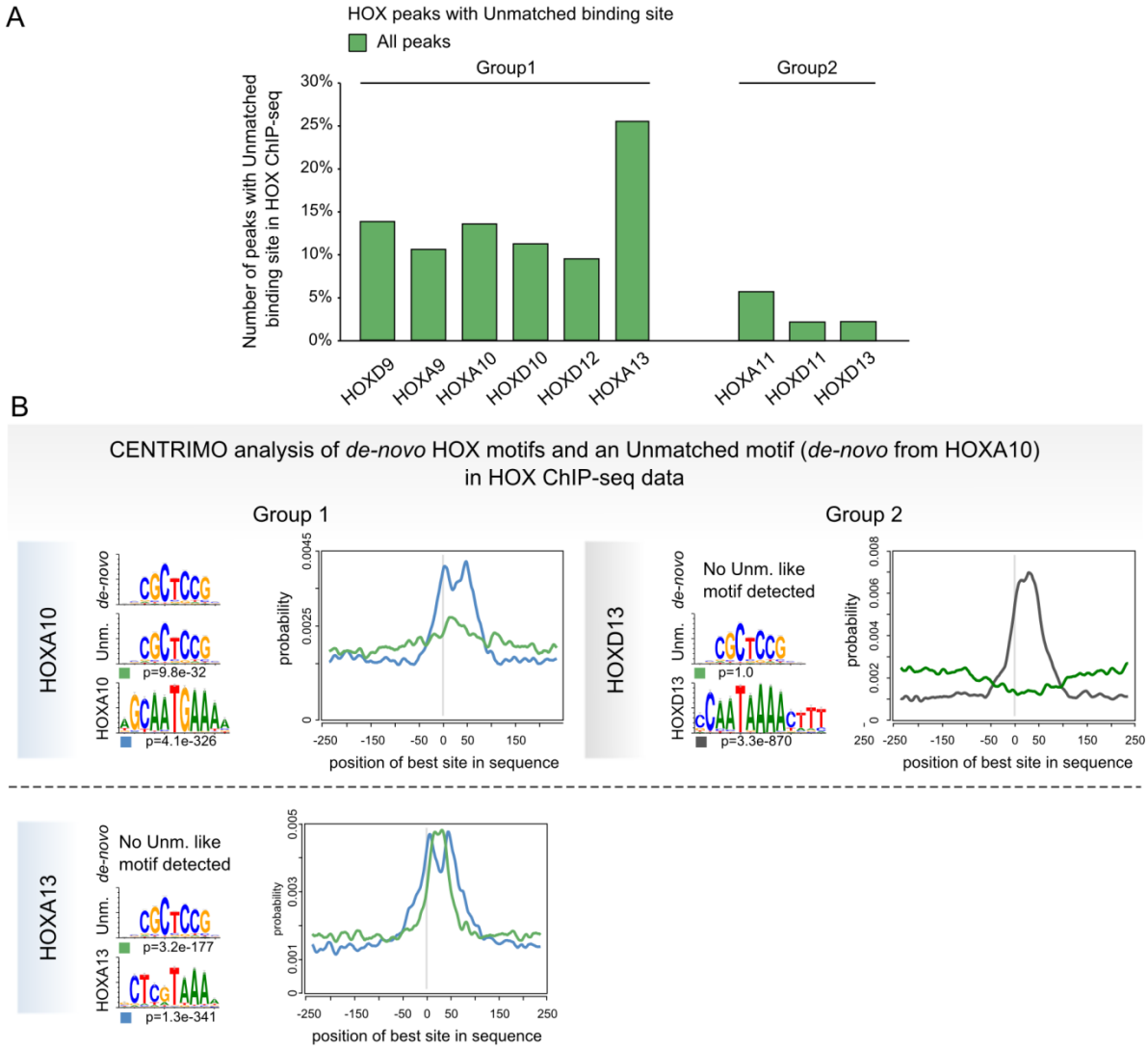


Figure 4.12 "Unmatched" motif is present mainly in the HOXA13 binding data as a secondary motif.

A) Quantification of the "Unmatched" motif in HOX data with FIMO tool (Grant, Bailey and Noble, 2011). B) Probability of finding the "Unmatched" motif and "Unmatched" motif positioning in the HOX binding data. Analzed with Centrimo tool (Bailey and MacHanick, 2012).

phenomenon. It will be interesting to see once this motif is associated with a protein how it will link to the HOXA13.

4.5.7 Quantitative and Qualitative Analysis of CTCF as a secondary Motif

Positioning and Abundance of the CTCF Motif in HOX-TF Binding Sites

Secondary motif analysis uncovered a *de novo* motif that was nearly a perfect match to the published CTCF motif. So far, CTCF has not been associated with HOX-TFs.

First, HOX-TF binding sites were examined for CTCF motif position and abundance. Centrimo and FIMO analyses demonstrated that CTCF motif is centrally enriched and abundant in Group 1 but not Group 2 binding sites (**Figure 4.13A** and **B**).

Second, identified differential CTCF abundance was examined in detail. It was important to make sure this not a background signal observed only in weakly enriched HOX sites. To account for this, peaks of all nine HOX-TFs were split into three groups according to their binding strength, top 1 000, top 10 000 and all peaks and then investigated for CTCF motif abundance. While the results were expectedly[5] mildly skewed towards all peaks, the overall differential distribution mainly stayed the same (**Figure 4.13A**). All peaks of Group 1 show a high abundance of the CTCF motif. Here, 22% of HOXA10 peaks carry a CTCF motif. This occupancy barely changes for the analysis of HOXA10 top 10 000 peaks. However, there is a more evident depletion of CTCF motif in HOXA10 strongest bound 1 000 peaks where CTCF is present at just 17% of the binding sites. Interestingly, the CTCF motif abundance in HOXA13 peaks does not change in three tested groups and is maintained at 23-24%. Conversely, Group 2 demonstrated consistently weak enrichment for CTCF motif in all three groups, ranging between 3-9%.

Altogether, the investigation of the secondary motifs uncovered three non-HOX motifs as a possible HOX cofactors. Out of these three motifs, only CTCF was highly abundant and centrally enriched at HOX binding sites. Moreover, CTCF abundance exhibited differential pattern between Group 1 and Group 2 binding sites, partially explaining novel subgrouping discovered in the PCA analysis.

[5] Since the strongest bound peaks are almost always direct peaks, they are enriched for the primary motifs. Therefore, the analysis of strongest peaks is skewed against secondary motifs, by design.

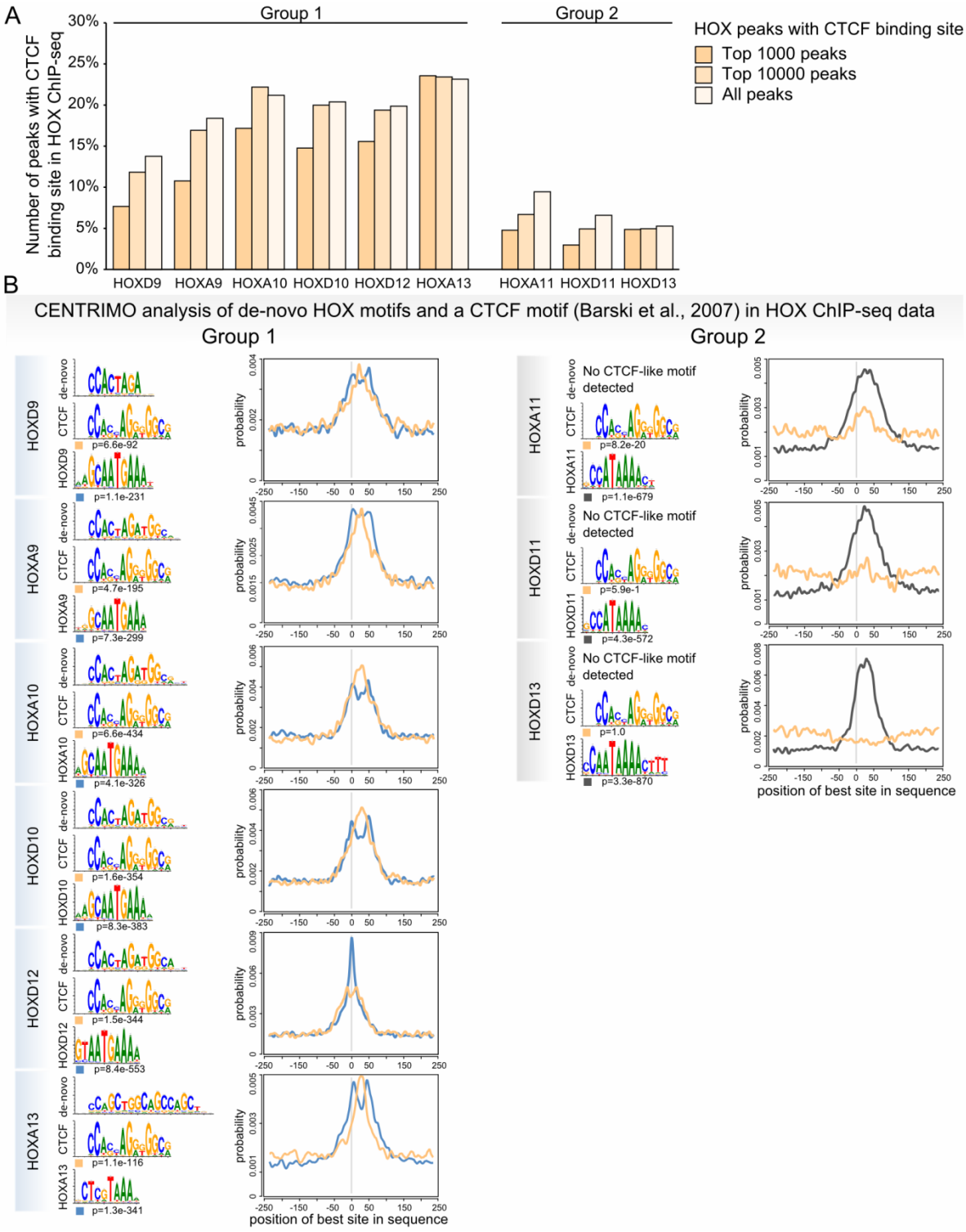


Figure 4.13 CTCF motif is more abundant in the Group 1 than in Group 2 HOX.
A) Quantification of the CTCF motif in HOX data with FIMO tool (Grant, Bailey and Noble, 2011). B) Probability of finding the CTCF motif and CTCF motif positioning in the HOX binding data. Analyzed with Centrimo tool (Bailey and MacHanick, 2012).

Delineation of HOX and CTCF Motifs in HOX-TF Binding Sites

De novo motif investigation and downstream analyses of primary and secondary motifs indicated a remarkable heterogeneity and complexity at HOX binding sites. Primary and secondary motifs can be present under same or different HOX binding sites. These two possibilities suggest different biological setups. Motifs present under the same binding site infer a possible co-binding and motifs present under different sites indicate likelihood of HOX being tethered to the DNA. Therefore, it was important to understand the relationship between HOX and CTCF motifs at HOX binding sites.

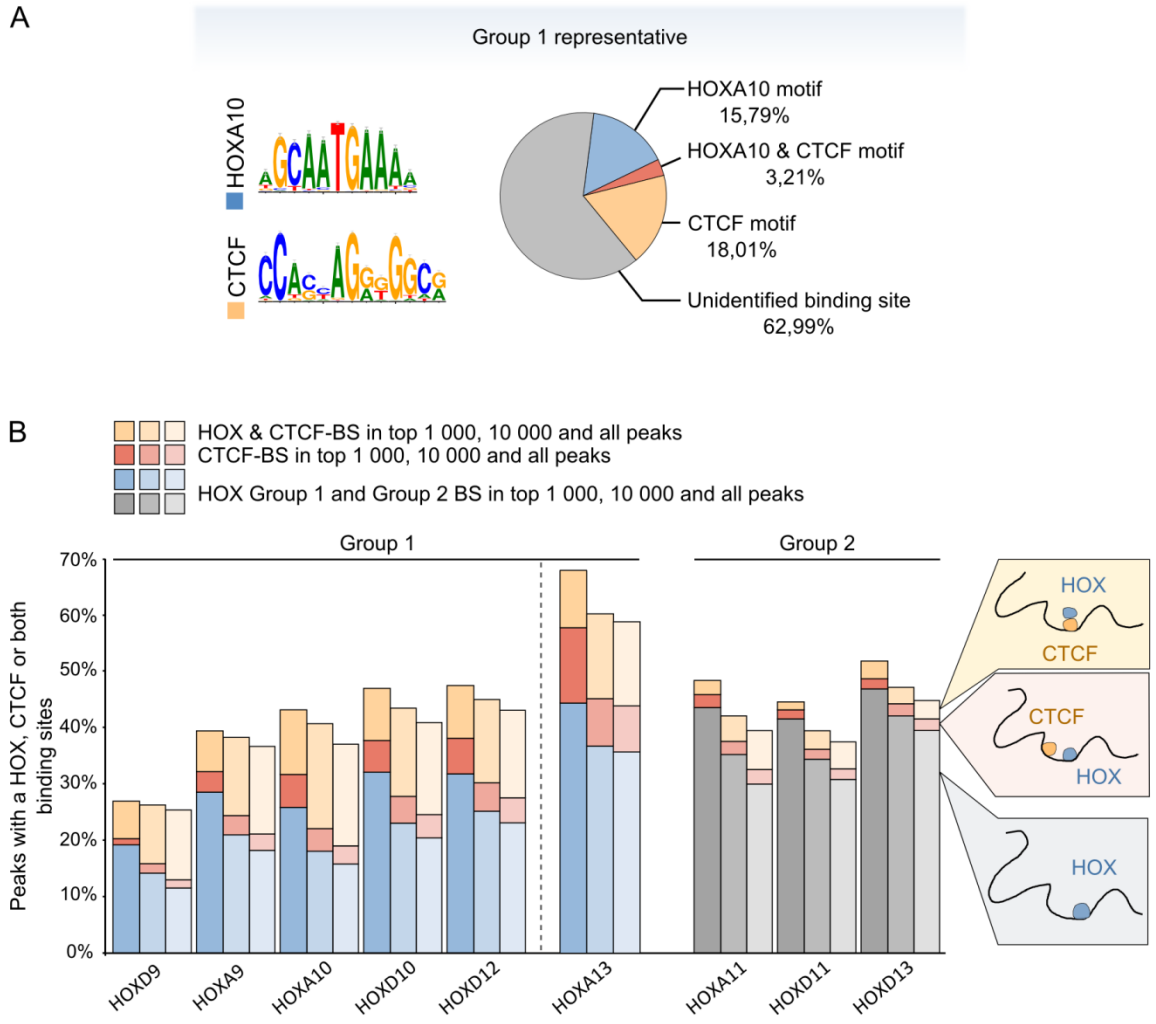


Figure 4.14 CTCF and HOX motifs are not found under the same HOX binding sites.
A) On the right: primary and secondary *de novo* motifs quantified and sorted according to presence of other motifs under the same binding site for the HOXA10 dataset. On the left: the motifs represented as discovered by the *de novo* motif analysis for HOXA10 and as published for CTCF (Jolma *et al.*, 2013). B) Primary and secondary *de novo* motifs quantified and sorted according to the presence of other motifs under the same binding site. Analysis was performed for all HOX top 1 000, top 10 000 and all bound peaks with FIMO (Grant, Bailey and Noble, 2011).

To approach this issue, using FIMO, I delineated HOX binding sites that contain only a CTCF, only a HOX primary motif, or both under the same peak. The presence of the both motifs underlying the same binding site is unexpected since HOXA10 and CTCF motifs are quite different (**Figure 4.14A**). Such situation would hint towards a cofactor-like relationship where the two TFs either, bind the DNA together or perhaps interact on a protein-protein level. Here, I again, subdivided all nine HOX binding profiles into the strongest 1 000, 10 000 and all peaks to make sure that observed effect is not a background effect. Then, I identified peaks carrying primary HOX motifs, CTCF motif, and checked if any of the peaks contained both motifs. This analysis showed that primary HOX motifs and the CTCF motif are most often present under different binding sites, regardless of peak enrichment. In most cases, under a single HOX peak there is either a HOX motif or a CTCF motif present, rarely both, except for HOXA13 TFBS where a significant fraction of the peaks carried both motifs (**Figure 4.14A** and **B**).

These results indicated that CTCF is a putative HOX co-factor and that it is likely helping to tether HOX to the DNA when there is no HOX motif.

4.6 Analysis of HOX Binding at Functional Chromatin

Functional Chromatin in *HOX* non-expressing Cells

To better understand the functionality TF binding sites it is helpful to identify at which functional chromatin marks the binding occurs. Unfortunately, not much chromatin modification data are available for chicken. Therefore, it was convenient that another study was conducted, concomitant with this one, in the same chMM system. This parallel study investigated

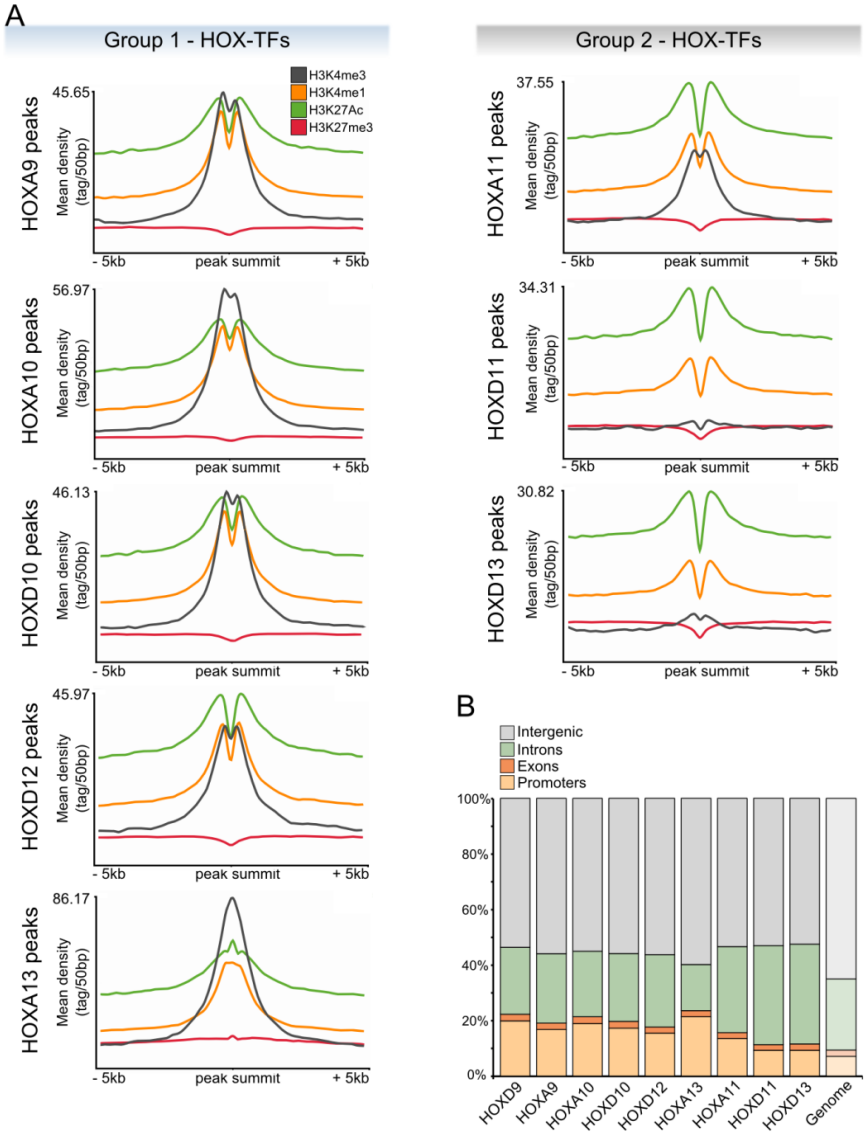


Figure 4.15 Group 1 HOX-TFs bind more often to H3K4me pre-marked regions.
A) seqMINER analysis of HOX binding sites at the marked functional chromatin. Active promoters (H3K4me3 and H3K27Ac); active enhancers (H3K27Ac and H3K4me1); poised enhancers (H3K4me1); repressed regions (H3K27me3); and bivalent regions (H3K4me3 and H3K27me3). B) HOX-TF binding at annotated genomic regions.

developmentally important chromatin marks at the day five of the chMM, uninfected cultures. These experiments were performed by Dr. Mickael Orgeur and published in the scope of his thesis and his manuscript in preparation (Orgeur, 2016). In this study, these data were used in order to differentially investigate position of HOX binding sites in respect to the functional chromatin.

Histone modifications were mapped for both, functionally active (H3K4me1, H3K4me3, and H3K27Ac), and repressive chromatin (H3K27me3). Then, I used seqMINER to investigate functional chromatin marks in a 10kb radius of the HOX-TF binding (Ye *et al.*, 2011). This analysis uncovered few general and Group-specific features of HOX-TF binding in respect to the functional chromatin. First, all HOX partially bind the regions enriched in functionally active chromatin related to enhancers (H3K4me1 and H3K27Ac), but not in functionally repressive chromatin (H3K27me3) (**Figure 4.15A**, **Figure 4.16A** and **B-**for the heatmap of H3K4me3 and H3K27me3 data, **Appendix 8-** for the heatmap of H3K27Ac and H3K4me1 data). Second, all HOX bind in a local minimum of the histone modification that is surrounded by the two local maxima on either side.[6] Active chromatin, promoter specific mark (H3K4me3) is differentially enriched between Group 1 and Group 2. More specifically, chromatin modifications marking poised/active (H3K4me1) and active (H3K27Ac) enhancers were present in all datasets while the mark characterizing active promoters (H3K4me3) was present only in vicinity of Group 1 and HOXA11 binding sites (**Figure 4.15A**).

Next, genomic positions of HOX peaks were examined to investigate genomic annotations of all peaks regardless of their functional chromatin status. Peaks could be located either in 1) promoters (-5 kb to +2 kb around the TSS), 2) exons, 3) introns, or 4) intergenic space (the rest of the genome). While most peaks mapped at intergenic regions, in total it was less than expected (**Figure 4.15B**). While the 65% of the genome is annotated as intergenic space between 52% (HOXA13) and 60% (HOXD13) of the peaks were found in these regions (**Figure 4.15B**). There was a slight increase in peaks located at annotated promoters. Here HOXD11/13 peaks mapped to 9% and HOXA13 to 21% at promoter regions, as compared to the 7% of the genome belonging to annotated promoters (**Figure 4.15B**). This finding recapitulated the tendency of Group1 to bind more often at the active promoters (H3K4me3) than the Group 2. Conversely, exon regions harboured only 2-2.5% of HOX peaks (**Figure 4.15B**). Lastly, together with the peaks in promoters, the biggest difference occurred with HOX binding at intronic regions.

[6] This binding pattern is expected as TFBS are centered at the nucleosome free region. Naturally, such nuclesome free regions are devoid of histones and histone marks.

Genome partition uncovered that with this custom annotation 26% of the genome belongs to the introns. However, HOX-TF peaks mapped to 17% (HOXA13) and up to 36% (HOXD13) at the intronic regions, deviating apart from the expected range (**Figure 4.15B**).

Functional Chromatin Changes in HOX Expressing Cells

Investigation of HOX binding to the genome previously marked with either active or repressive mark shed light onto the functionality of these binding events. Furthermore, these results pointed out the differential preference of two groups with respect to the functional chromatin. These findings were used as a start point in the subsequent analysis where I monitored functional chromatin changes upon HOX overexpression. For this purpose, chMM cultures overexpressing HOXA10 and HOXD13 were used as representatives for the Group 1 and Group 2. ChIP-seq was performed as reported in **Chapter 3.6**. This time, however, two selected histone modifications were examined in these HOX overexpressing cultures, H3K4me3, and H3K27me3. H3K4me3 was selected due to the differential coverage between Group 1 and Group 2. Furthermore, H3K4me3 and H3K27me3 have a delicate balance over the HoxD cluster, and it has been shown that presence of HOX13 proteins can influence this balance (Beccari *et al.*, 2016; Sheth *et al.*, 2016). Furthermore, spreading of both marks is reportedly blocked by CTCF binding (Cuddapah *et al.*, 2009; Varun Narendra *et al.*, 2015). Therefore, I set out to examine changes in these two marks in overexpressions and control conditions.

To be able to examine functional chromatin changes over the CTCF sites, a CTCF ChIP-seq was performed. The analysis of the CTCF reproducibility and peak calling was performed as for all other HOX ChIP-seq (**Appendix 5**). Data were clustered with seqMINER using same parameters around HOXA10, HOXD13 or CTCF peaks (**Figure 4.16A, B,** and **C**). First, analysis was focused on the changes in chromatin marks centered at HOXA10 and HOXD13 peaks in their respective chMM cultures. Generally, coverage was higher and signal stronger in the data from the control ChIP-seq, making the control data always looks a bit noisier and with somewhat more pronounced signal. Overall, there is little change in the H3K4me3 and H3K27me3 mark distributions around HOXA10 and HOXD13 peaks in the control or HOXA10/HOXD13 cultures (**Figure 4.16A** and **B**). Furthermore, two histone marks (H3K4me and H3K27me) were largely devoid of signal around the HOXA10, HOXD13, and CTCF binding sites, especially in HOXD13 overexpression (**Figure 4.16A**, **B,** and **C**). This indicates that HOXD13 doesn't often bind to the pre-marked chromatin, at least with respect to these two marks. Apparent gain of H3K4me3 mark in the Cluster 2 and 3 indicates that a subset of

promoters in the HOXA10 overexpressing culture get activated. However, since the signal is stronger in the control culture[7], it is difficult to establish to which extent. Due to this, cluster 2 and cluster 3 were not the focus of subsequent analysis.

Next, I examined the change of H3K4me3 and H3K27me3 marks centered at the CTCF peaks that were identified from a CTCF ChIP-seq performed in the same system. This analysis, expectedly, showed that higher H3K4me3 signal at the CTCF sites in these data, as CTCF is known to bind at promoter regions (**Figure 4.16C**). More interestingly, it also showed there is an increase of the H3K27me3 mark over the subset of CTCF binding sites, both, in HOXA10 and HOXD13 overexpressing cultures (**Figure 4.16C**). These sites were then extracted and checked for their genomic position. Surprisingly, there was a gain of H3K4me3 mark at many promoters in cluster 4 (**Figure 4.16D**). In this cluster, 41% of CTCF binding sites mapped at promoters in comparison to 22% all CTCF binding sites. Furthermore, these binding sites were mostly devoid of any intronic binding. Since these binding sites located at many promoters I investigated GO terms association with these promoters as they are likely to direct targets of this gain of methylation. Interestingly, these promoters were enriched for GO terms associated with metabolic processes, specifically with RNA and DNA processing, cofactors, histone binding, and methylated residue binding.

[7] The signal was also monitored on an enrichment graph in seqMINER. While the signal was weaker in the HOXA10 chMM it was still present.

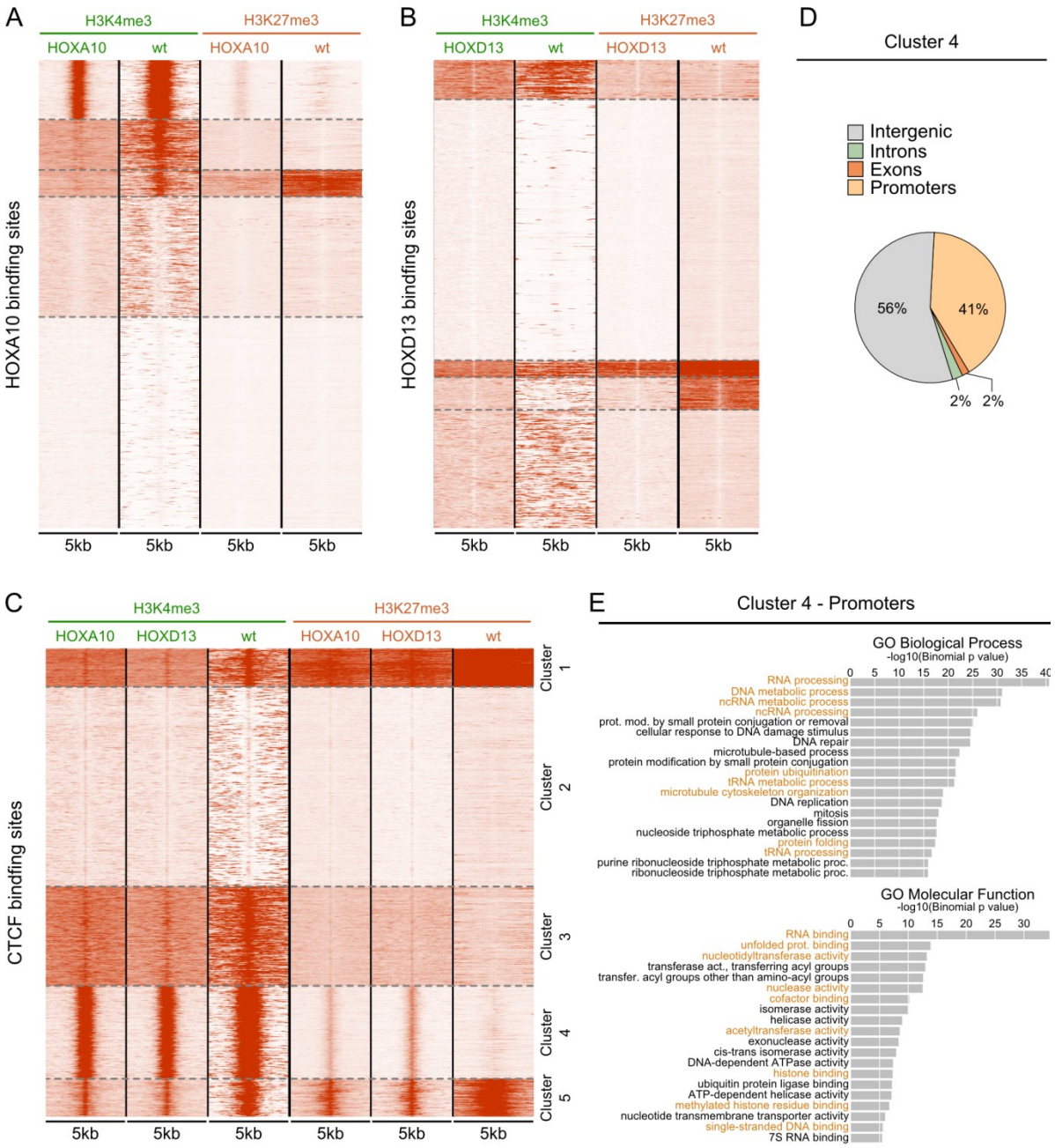


Figure 4.16 HOXA10 and HOXD13 expressing cultures promote a gain in H3K27me3 at a subset of CTCF binding sites.

A) seqMINER analysis of the H3K4me3 and H3K27me3 mark in the uninfected chMM culture and in the HOXA10 infected culture. The histone marks are centered at 5kb window around the HOXA10 binding sites B) seqMINER analysis of the H3K4me3 and H3K27me3 mark in the uninfected chMM culture and in the HOXD13 infected culture. The histone marks are centered at 5kb window around the HOXD13 binding sites. C) seqMINER analysis of the H3K4me3 and H3K27me3 mark in the uninfected chMM culture, HOXA10, and HOXD13 infected culture. The histone marks are centered at 5kb window around the CTCF binding sites. D) Genomic annotation the subset of CTCF peaks from the Cluster 4 in C). E) GO analysis of the promoters found as annotated in the Cluster 4.

4.7 Analysis of CTCF, RAD21 and HOX Co-binding

Co-binding Quantification

To corroborate previous results, it was essential to comprehensively and thoroughly investigate a possibility of CTCF-HOX co-binding. Earlier analyses detected and described CTCF motif at the HOX binding sites. However, so far these sites represent only putative co-binding, pending validation. To address this question, used CTCF ChIP-seq from same chMM setup, and then I investigated the binding of CTCF and its relation to HOX binding.

First, 22 357 CTCF binding sites were reproducibly identified (**Figure 4.17A**). Then, HOX peaks

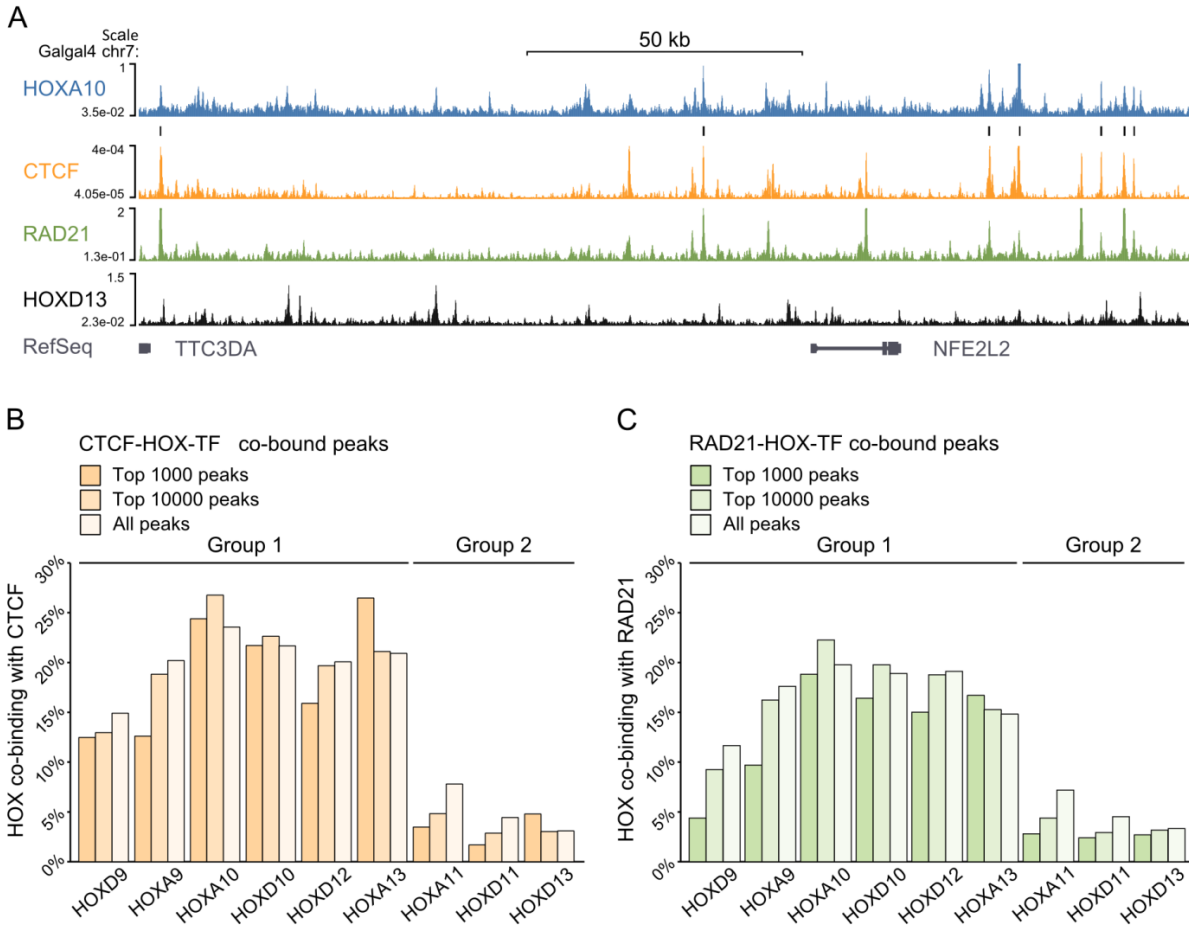


Figure 4.17 HOX Group 1 bind to a proportion of same genomic locations like CTCF and RAD21.
A) UCSC snapshot of the HOXA10, CTCF, RAD21 and HOXD13 binding in the Galga4 genome. B) Proportion of the binding sites that are occupying the same genomic locations in HOX and CTCF binding profiles for the top 1 000, top 10 000 and all bound peaks. C) Proportion of the binding sites that are occupying the same genomic locations in HOX and RAD21 binding profiles for the top 1 000, top 10 000 and all bound peaks

were once again subdivided into strongest bound 1 000, 10 000 and all peaks and overlap between HOX peaks and CTCF peaks was investigated (**Figure 4.17B**). This analysis once again recreated, above defined, discrepancy between Group 1 and Group 2. Group 1 but not Group 2, frequently co-occupied same sites like CTCF. Co-binding in Group 1-CTCF ranged from 13% for HOXA/D9 to 27% for HOXA13 in the strongest bound 1 000 peaks. This changed to 15% and 24% for HOXD9 and HOXA10 in all bound peaks, respectively. Interestingly, just like in the examination of putative CTCF binding sites in HOX binding data, HOXA13 exhibits highest binding co-localization with CTCF in the most strongly bound 1 000 peaks. Furthermore, HOXA10 and HOXD10 co-bind same regions like CTCF equally abundant, regardless of peak strength (**Figure 4.17B**). Conversely, Group 2 rarely binds the same regions like CTCF. Co-binding ranges from 2% for HOXD11 and 5% for HOXD13 for the strongest bound 1 000 peaks. In all peaks, this changes to 3% and 8% for HOXD13 and HOXA11, respectively. These results corroborate earlier defined discrepancy between Group 1 and Group 2 in respect to co-binding with CTCF.

Second, to assess if HOX and CTCF co-bound sites are also occupied by Cohesin complex, I performed RAD21 (Cohesin subunit) ChIP-seq. Here, Cohesin was of particular interest as it is known to direct chromatin looping when co-bound with CTCF (Nora *et al.*, 2016; Schwarzer *et al.*, 2016). Upon initial quality analysis, 17 585 reproducible peaks were detected for RAD21 in chMM. All initial quality and reproducibility analysis were performed as for the rest of the TF ChIP-seq used here (**Appendix 6**). Then, HOX and RAD21 co-binding was investigated as previously for the CTCF. Overall, same pattern of discrepancy between Group 1 and Group 2 was uncovered; where Group 1 and Cohesin appear to bind same regions more often than Group 2 and Cohesin. This is also in line with above-discussed CTCF-HOX co-binding observation. In the strongest 1 000 peaks, Group 1 HOX share between 4% and 19% binding sites with Cohesin for HOXD9 and HOXA10, respectively (**Figure 4.17C**). In all peaks, the co-binding changes to 12% and 20% for HOXD9 and HOXA10, respectively (**Figure 4.17C**). Conversely, Group 2 HOX co-bind with Cohesin quite rarely. In the strongest bound peaks at best only 3% of HOXD13 co-binds with Cohesin. When compared to co-binding in all peaks, this changes to the maximum 7% co-binding for HOXA11 and Cohesin, indicating that reduced co-binding is a general feature of Group 2 regardless of peak strength (**Figure 4.17C**).

Characterization of Co-bound Sites by Motifs and Genomic Annotation

As HOX-CTCF and HOX-Cohesin co-binding display similar pattern of discrepancy between Group 1 and Group 2, I hypothesized that the origin of this similar pattern are triple bound HOX-CTCF-Cohesin sites. To approach this issue, first, I studied the overlap of all CTCF and RAD21 binding sites. In accordance with other studies[8], I found that 52% of all CTCF peaks were also bound by RAD21, and 66% of the all RAD21 peaks were also bound by the CTCF (Andrey *et al.*, 2017). Next, triple HOX-CTCF-Cohesin occupied peaks were examined. Here, Group 1 exhibited remarkably frequent triple binding. In fact, HOX-CTCF sites were more frequently co-occupied by Cohesin than all genome-wide CTCF binding sites (**Figure 4.18A** and **Appendix 7**). Same trend was present in the Group 2 data but to a lesser degree (**Figure 4.18A** and **Appendix 7**).

To identify the binding mode of co-bound bound sites, I quantified HOX primary and CTCF motifs in co-bound peaks. This analysis showed a remarkable shift from primary HOX motifs

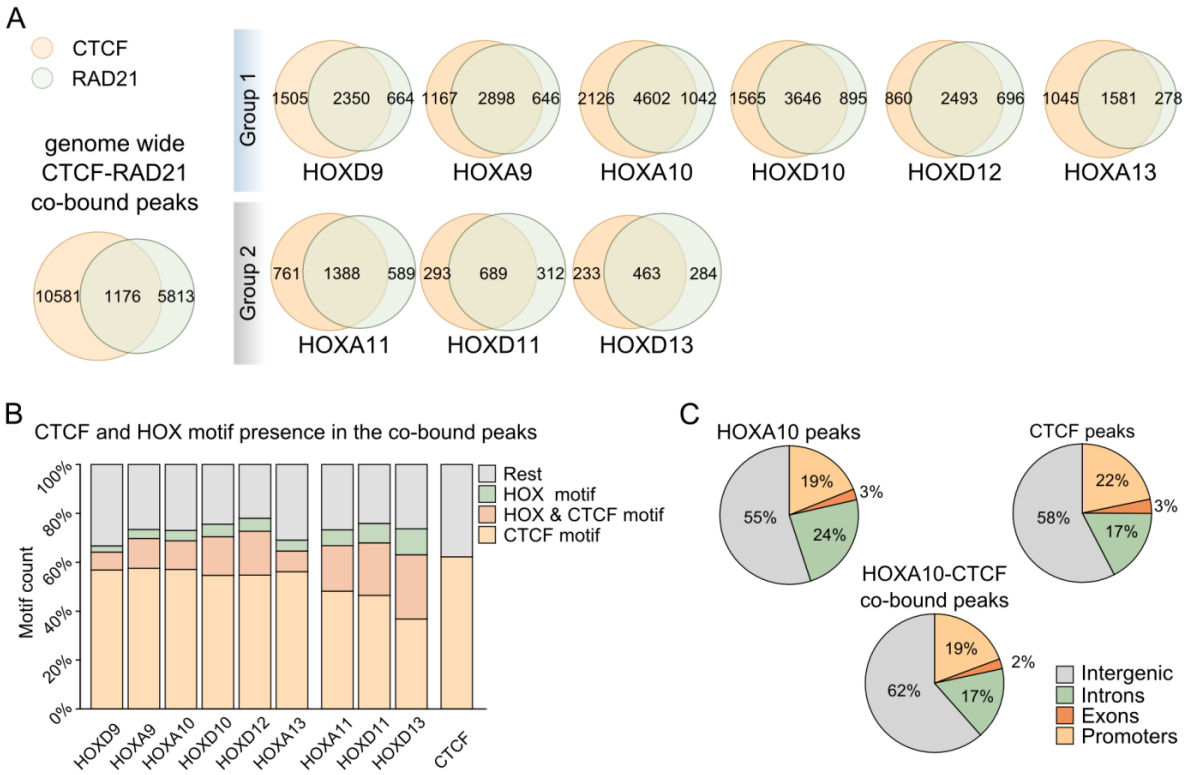


Figure 4.18 HOX-CTCF bound genomic sites often contain the CTCF motif but not HOX motif.

A) HOX-CTCF bound genomic sites investigated for the presence of the binding of the RAD21. B) Quantification of the CTCF and primary HOX motifs in the CTCF-HOX bound genomic sites by FIMO tool (Grant, Bailey and Noble, 2011). C) Annotation of the genomic locations of the HOXA10, CTCF and HOXA10-CTCF binding sites.

towards CTCF motif (**Figure 4.18B**). HOX-CTCF sites largely depend on CTCF motif as a DNA binding mode as this motif is present in up to 73% of the co-bound sites. In comparison, only 62% of the all CTCF peaks carry a CTCF motif (**Figure 4.18B**). Furthermore, HOX-CTCF co-bound sites showed a mild discrepancy between the Group 1 and 2 in the content of the binding sites. Here again, Group 2 contains more primary HOX and CTCF motifs underlying the same binding sites, than it is the case for the Group 1 (**Figure 4.18A**).

Lastly, genomic annotations of the co-bound peaks were examined to determine whether a subset of co-bound peaks have a preference for a certain genomic location and/or function (e.g. promoter regions). In comparison to the genomic locations of the entire HOX binding sites, there is aminor increase in the intergenic regions (**Figure 4.18C** and **Appendix 9**). Overall very little change is present in the genomic annotation of co-bound peaks in comparison to total HOX peaks. Therefore, CTCF-HOX co-occupancy is not preferentially occurring in any genomic location here examined indicating that majority of binding happens outside promoters, likely at *cis*-regulatory elements or at regions important for local genome architecture.

4.7.1 Demonstration and Delineation of HOX-CTCF Protein-protein Interaction

Genomic locations co-occupied by CTCF and HOX mostly rely on CTCF-DNA interactions. However, the evidence of how they could co-bind these regions is still very circumstantial. Therefore, I set out to demonstrate protein-protein interaction between HOX and CTCF. For this purpose, Proximity Ligation Assay (PLA) was employed. PLA is an essay that gives a distinct, dotted signal under the microscope when two proteins are in extreme proximity (i.e. when they are in a complex) (**Figure 4.19A**). Furthermore, PLA is an *in situ* and quantifiable assay. These properties allow additional insights into the abundance of HOX-CTCF interaction in individual nuclei since co-binding occurs in, both, Group 1 and Group 2 but to a different degree.

To investigate the potential of Group 1 and Group 2 to co-bind with CTCF, PLA assay was performed with three protein combinations: 1) HOXA10-CTCF, 2) HOXD13-CTCF, and 3) RAD21-CTCF (see **Chapter 3.5**). HOXA10 was a representative for the Group 1, HOXD13 for Group 2 and RAD21 a positive control. As negative control, cells not expressing HOX proteins were used. Signal was present in the nuclei of positive control, HOXA10, and HOXD13 expressing cells, but not in the negative control. Closer inspection of the cell nuclei, in accordance with the ChIP-seq results, uncovered that signal in the between CTCF and HOXD13

was weaker than in the positive control or CTCF-HOXA10 sample (**Figure 4.19B**). Furthermore, upon signal quantification, it became evident that there is significantly less signal in HOXD13 than in the HOXA10 sample (**Figure 4.19C**).

Finally, to decipher how HOXA0 and CTCF interact I set out to determine which part of the HOXA10 protein is responsible for the contact with CTCF. If the responsible domain was anything other than Homeodomain, it would be possible to investigate HOX binding in the absence of a CTCF interacting domain and help directly determine the causality of HOX-CTCF co-binding. For this purpose, seven HOXA10 deletion constructs were designed and cloned in frame with FLAG-tag into the RCASBP(A) vector. Deletion constructs were named HOXA10-

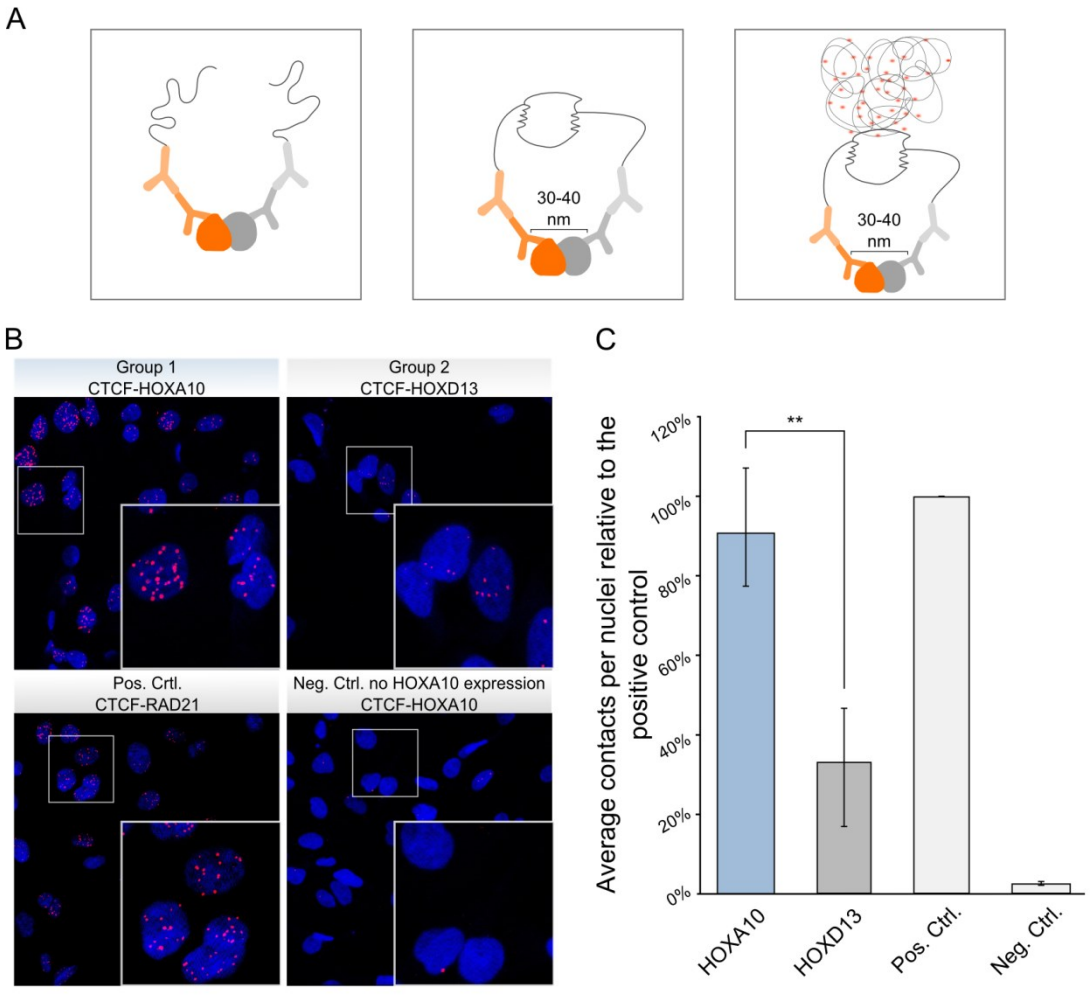


Figure 4.19 HOXA10 and HOXD13 interact with the CTCF.
A) A schematic overview of the Proximity Ligation Assay (PLA). B) PLA visualized under the confocal microscope for the HOXA10-CTCF, HOXD13-CTCF, CTCF-RAD21 (positive control) and CTCF negative control) signal. C) Quantification of the PLA signal detected in B). The assay was repeated in three different biological replicate. Student's T-test was employed to quantify the significance of the contact frequency between HOXA10 and HOXD13 (Student's T pval<0.005).

Δ1-HOXA10-Δ7. First five deletion constructs cover HOXA10 protein in sequential deletions, and the last two were based on predictions of the possible protein interaction sites (**Figure 4.20A**). Deletion constructs were first tested to make sure that they are localized in the nuclei. Naturally, at least one construct was expected to not have nuclear localization as the nuclear localization signal (NLS) is supposed to be deleted. These experiments showed that HOXA10-Δ5 is largely absent from the nuclei, indicating that the NLS is located in the C-terminal part of the HOXA10. Next, I performed the co-IP and PLA to test HOXA10 deletion constructs and CTCF for protein-protein interactions. Co-IP results from the nuclear lysate were reproducible and demonstrated loss of contact between CTCF and HOXA10-Δ5 (**Figure 4.20B**). Importantly, the HOXA10-Δ5 carries Homeodomain and the NLS signal. Therefore it was impossible to know for sure if this part of the HOXA10 protein is responsible for the contact with the CTCF.

PLA results, however, yielded a more complex picture of HOXA10-CTCF interaction. Indeed, the most extreme loss of signal appeared when testing CTCF and HOXA10-Δ5 interaction (**Figure 4.20C**). Since PLA is a quantitative assay, I was able to measure and compare the potential of other deletion constructs to interact with CTCF as well. Deletion constructs Δ2, Δ3, Δ6, and Δ7 all displayed a mild reduction in the contact frequency in comparison to the full-length HOXA10 protein. This observation can be explained in at least two plausible ways. First, the construct deletions could destabilize HOXA10 protein causing reduced contact frequency. Second, identified HOXA10-CTCF interaction is not a direct interaction but rather a part of the larger complex. In the complex, deletions of different HOXA10 protein parts could impact the complex stability and cause the observed reduction in contact frequency.

Altogether, these experiments demonstrated that HOXA10 and CTCF interact on a protein level. However, it was impossible to unambiguously define the region of the HOXA10 protein responsible for the interaction with CTCF and whether HOXA10 and CTCF directly interact or are a part of a bigger complex where several HOXA10 domains establish and/or maintain the protein-protein interactions.

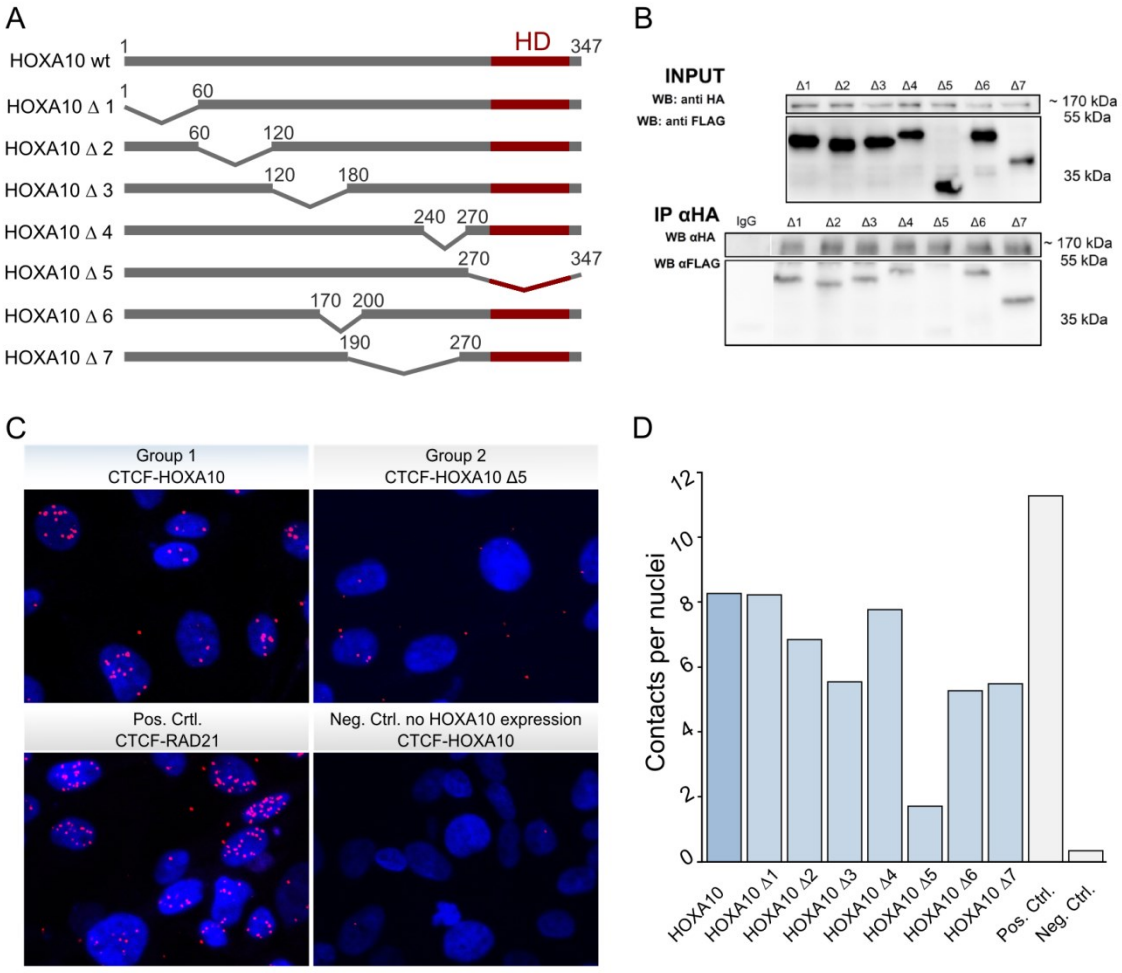


Figure 4.20 HOXA10 likely contacts CTCF indirectly.
A) Schematic representation of the HOXA10 deletion constructs. B) A co-immunoprecipitation (co-IP) assay performed with the HOXA10 deletion constructs and the CTCF. C) Visualized PLA assay with the HOXA10 deletion constructs HOXA10 wt-CTCF, HOXA10 Δ5-CTCF, CTCF-RAD21 (positive control) and CTCF only (negative control). D) Quantification of the signal for all the HOXA10 Δ-CTCF pairs, in exemplary showed in C). This experiment was performed only once with multiple sections counted and averaged the contact per nuclei.

5 Discussion

5.1 Functional Redundancy in the Induced Regulatory Programs

5.1.1 Redundancy beyond Paralogy Groups

Hierarchical clustering of the strongest differentially regulated genes in the chMM uncovered an unusual pattern. Firstly, the HOXA13 induced regulatory program clustered far from all other tested HOX induced regulatory programs, including its paralogue HOXD13. This is somehow in contradiction with the functional redundancy found during digit morphogenesis (Fromental-Ramain, Warot, Messadecq, *et al.*, 1996). However, it is likely that only a subset of downstream target genes are necessary to conduct the HOX13-induced digit morphogenesis and that they are out shadowed by a large number of differentially regulated genes. Secondly, HOXD12 and HOXD13 induce highly similar regulatory programs, separated from the ones of HOXD9 and HOX10/11. This finding is in line with the loss-of-function and deletion experiments in mice. There, a loss-of-function mutation in the *Hoxd13* gene causes severe polydactyly and delay in ossification (Kmita *et al.*, 2002; M Kmita *et al.*, 2005). However, the deletion of the *Hoxd13* has a significantly different functional outcome. Upon *Hoxd13* deletion, *Hoxd12* moves to the *Hoxd13* genomic position with respect to digit regulatory regions. The change in genomic position causes *Hoxd12* to adopt a *Hoxd13* wild-type expression domain. This regulatory compensation is further complemented by a functional compensation, as these mice are not affected by polydactyly like the *Hoxd13* loss-of-function mice. The functional compensation is a direct evidence of redundancy between *Hoxd13* and *Hoxd12*. This genetic data is further validated by the close clustering of HOXD12 and HOXD13 regulatory programs described in this study.

Interestingly, a similar experiment was conducted to investigate the regulatory and functional compensation between *Hoxd11* and *Hoxd13*. There, *Hoxd13* and *Hoxd12* deletion caused *Hoxd11* to take over the genomic position of the last *Hox* gene in the cluster. Like in the example before, this change triggers *Hoxd11* to overtake *Hoxd13* expression pattern. However, this mutant cannot rescue the *Hoxd13* loss-of-function phenotype as *Hoxd12* can. Therefore, not all neighboring genes are equally functionally redundant and interchangeable (Kmita *et al.*, 2002; Marie Kmita *et al.*, 2005; J Zakany and Duboule, 2007). These findings go well in line with clustering of induced regulatory programs, where indeed transcriptional programs of HOXD12 and HOXD13 resemble more than the HOX10/11 and HOXD9.

5.1.2 Impact of Evolutionary Adaptation on Differentially Induced Regulatory Programs

HOX-induced regulatory programs are in general very similar, especially for above-discussed groups. However, PG9 and PG13 paralogy groups induce highly specific regulatory programs. This is in sharp contrast with other paralogy groups, such as PG10 or PG11, which are strikingly redundant. I hypothesize that this is the result of a higher gene specialization for PG9 and PG13 groups during evolution. This is linked to the fact that both paralogy groups, in contrast to others, still retain all four copies of ancestral genes. This copy number effect might have reduced the evolutionary pressure on each copy and allowed for gene neofunctionalization (Gehring, Kloter and Suga, 2009). Therefore, PG9 and PG13 genes had more opportunity to diverge throughout the time, which caused some paralogues to obtain new functions while others retained the old ones.

5.1.3 HOX Autoregulation

Since HOX-TFs induce similar regulatory programs in the chMM, it was imperative to understand if there is any HOX autoregulation present in these cultures. So far, there is little direct evidence for autoregulation in *trans* and in particular between different clusters.

In this dataset, HOX autoregulation was detected in all nine HOX induced programs, both within the same cluster and between the clusters. Interestingly, the pattern of autoregulation appears to have a posterior bias for all HOX proteins, except for HOXA9. Specifically, overexpression of HOX9-12 TFs induces upregulation of *HOXA13* and *HOXD13* genes. This posterior bias is a fascinating pattern of autoregulation that is somewhat reminiscent of the posterior prevalence model of HOX regulation (Duboule, 1994; Duboule and Morata, 1994), in which the most posterior *HOX* expressed in tissue always determines the tissue identity. Furthermore, this pattern of HOXA13 and HOXD13 upregulation is comparable to the sequential opening and progressive transcriptional activation of the HoxD cluster. There, progressive cluster opening upregulates genes sequentially which helps upregulate the expression of their posterior neighbor until it reaches the HOX13 paralogues. Upon expression of HOX13 paralogues, a switch from proximal to distal patterning occurs and expression terminates (Beccari *et al.*, 2016). This switch is essential for the proper patterning of the distal limb.

Altogether, the data indicates that there could be an additional intrinsic HOX mechanism to activate most posterior genes, ensure proper development and termination of HOX expression.

5.1.4 Developmental Context of Transcriptional Redundancy

Above discussed novelties need to be put in a proper developmental context to be fully and correctly understood. In vertebrates, budding of the mesenchymal limb is influenced by different morphogen concentrations and signaling pathways building the "ground plan". When any group of *Hox* genes are expressed in cells with the same "ground plan", they likely impact their targets in a paralogue specific manner. However, if *Hox* genes are expressed under a different "ground plan", it is likely that the target gene is a unique target (Mann, Lelli and Joshi, 2009; Wolpert, Tickle and Martinez, 2015). This might seem at odds with the finding that HOX proteins induce many similar regulatory pathways in chMM. However, this merely means that they possess a potential to induce these programs in an *in vivo* situation. Moreover, *Hox* genes are systematically expressed in a combinatorial fashion *in vivo*, which might trigger particular transcriptional outcomes that could not be quantified in the present study. So far, it is unclear how the regulatory regions bound by HOX-TFs function and if they need the activity of signaling pathways to be potentiated/activated in order to conduct their transcriptional effect properly. The "ground plan" that is slightly different from one part of the limb to another can, thus, potentially drive HOX-TFs to affect their targets differently, depending on the positional identity.

During vertebrate limb bud development, three main signaling centers are delivering signaling molecules of WNT, FGF, SHH, and RA (Wolpert, Tickle and Martinez, 2015). Together, these signaling pathways determine positional identity of cells in the developing limb bud. chMM, however, does not have any signaling molecules present. Therefore, the transcriptional readout of the chMM-induced programs might not represent accurately the *in vivo HOX* target gene specificities. Finally, while the chMM system allows one to study the binding quite accurately, it is not trivial to interpret the transcriptional readout in the *in vivo* context.

5.2 Understanding Discrepancy between HOX Binding and Target Regulation

5.2.1 Reproducible Low-affinity HOX Binding Sites

ChIP-seq is a technique that enables mapping of the protein-DNA interactions. Therefore, due to the nature of these interactions, accurate representation of binding profiles primarily depends on antibody quality, abundance and stability of the protein. Here, HOX-TF ChIP-seq data have been generated in identical conditions, from the same pool of cells, and using the same antibody. Nevertheless, HOX binding profiles displayed constant and reproducible differences in the enrichment. More specifically, the more centrally positioned HOX, PG9, and PG10, had more peaks of lower enrichment than other HOX proteins. Importantly, PG9 and PG10 binding profiles were reproducible in at least three independent biological replicates. Thus, low enrichment could not be attributed to technical issues, but rather to biological properties of HOX-TFs.

HOX-TFs are notorious for their dependency on cofactors, most notably TALE proteins. Multiple studies in the past two decades have been investigating the ability of TALE proteins to influence direct HOX-DNA binding (Mann, Lelli and Joshi, 2009; Crocker *et al.*, 2015; Merabet and Lohmann, 2015; Merabet and Mann, 2016). Specifically, Crocker et al. (2015) showed that HOX-TFs co-bind with TALE proteins to low affinity *bona fide* HOX sites to ensure proper target regulation. There, Ultrabithorax (Ubx) binds with Extradenticle (Exd) to low-affinity binding sites at the *shavenbaby* enhancer. Additionally, multiple low-affinity sites are needed to establish robust expression of *shavenbaby*. This is in agreement with Farley et al. (2015), where they found a cluster of sub-optimal (low affinity) transcription factor binding sites are required for proper expression of a developmental gene. Importantly, if only high-affinity binding sites are present, the target gene is grossly miss- and overexpressed. Therefore, low-affinity binding sites, not only allow binding of developmental TFs but, are essential for the proper gene expression and embryonic development of an organism.

Indeed, while binding of PG13 groups showed more direct and monomer-like binding, more centrally positioned PG9 and PG10 did not. Specifically, *de novo* motif discovery from PG13 uncovered primary motifs that are almost unchanged from Homeodomain-DNA binding motifs described *in vitro*. However, other HOX proteins exhibited considerably changed motifs in comparison to their *in vitro* counterparts (M F Berger *et al.*, 2008; Jolma *et al.*, 2013). These

findings resemble HOX-TALE co-binding, which targets low-affinity sites with changed motifs. Furthermore, anterior and central HOX-TFs are known to depend more on co-binding with TALE proteins than very posterior HOX-TFs (Mann, Lelli and Joshi, 2009). Together, these findings suggest the very posterior HOX-TFs depend more on direct, perhaps monomer binding, than more central PGs. It is possible that this is driven by the divergence in PG13 protein sequences along with the maintenance of complete four paralogs (see **Chapter 5.2.3**). Finally, these findings also indicate that PG9 and PG10 low-enrichment sites are likely a HOX-cofactor bound low-affinity sites, or perhaps sites where HOX proteins are tethered by other proteins. Both possibilities determine these sites a biologically valuable contribution to the study of HOX binding.

5.2.2 Common HOX binding sites

Posterior HOX-TFs bind many same genomic locations throughout the genome. Specifically, up to 86% (for PG11) of the binding sites are shared within paralogy groups and up to 71% (for HOXA9-HOXD10) between different paralogy groups. Evidently, HOX proteins exhibit a remarkable binding redundancy. However, the binding redundancy does not directly translate to identical target gene response. In *Drosophila*, Distalless (*dll*) gene drives leg development in the thoracic segment and is bound by Ubx, AbdA, AbdB, Scr, and Antp. On the other hand, in the abdominal segment, *dll* is repressed by the Ubx, AbdA, and AbdB preventing leg development. Therefore, it is clear that the *dll* is a *bona fide* target of many HOX-TFs. Depending on the segment identity and presence of cofactors (co-repressors, co-regulators) and/or signaling molecules, bound target response will be modulated, either repressed or activated (see **Chapter 5.2.3**). Thus, HOX binding doesn't always need to be specific, as there is an additional layer of specificity achieved by tissue identity, cofactors, and signaling molecules (Mann, Lelli and Joshi, 2009; Merabet and Mann, 2016). Furthermore, it is necessary to keep in mind that *in vivo HOX* genes always come in combinations, which further complicates the relationship between binding and target regulation. This relationship can be either collaborative (Hox code theory) or more antagonistic (Posterior prevalence theory) (Kessel and Gruss, 1990; Duboule and Morata, 1994). Since evidence from literature and this study don't exclude either of these theories, it is likely that potential collaboration or antagonism is too, governed by the timing and tissue identity.

5.2.3 Direct HOX-DNA Binding Specificity

TFs usually rely on the direct binding. However, additional cofactors can enhance and/or change the binding specificity by either tweaking the direct binding specificity or by driving the indirect binding (Slattery *et al.*, 2014). HOX proteins are known for their discrepancy between their binding and target regulation specificity, known as a Hox paradox (Mann, Lelli and Joshi, 2009). One layer of this paradox is due to the limited knowledge of the HOX-DNA binding mechanism. So far, the best known HOX cofactors are proteins from the TALE family. TALE proteins bind the HOX-TFs on either side of the Homeodomain and change the specificity of HOX-DNA contact (Merabet and Mann, 2016). Several large scale *in vitro* studies have been conducted in the past ten years attempting to decipher HOX-DNA binding in different animals (M F Berger *et al.*, 2008; Jolma *et al.*, 2013). However, due to the large-scale approach, these studies mainly focused on Homeodomain-DNA binding, therefore providing a high affinity monomer-like binding motif. This posed a difficulty, when interpreting HOX binding, since it is known that they rarely bind as monomers. Later, Slattery et al. (2011) and Jolma et al. (2015) expanded their experimental design to study HOX-cofactor specificity as well. While these experiments provided an invaluable source of HOX binding data, they were conducted *in vitro* and with a narrow selection of cofactors. In comparison, this study for the first time provided evidence of *in vivo* vertebrate HOX binding. Therefore, it was imperative to compare the discovered HOX motifs with known Homeodomain and HOX motifs.

Primary motifs identified here show striking discrepancy from either Homeodomain-DNA or monomer HOX-DNA binding (Berger et al. 2008; Jolma et al. 2013). Profound differences in motifs are likely due to the presence of full-length proteins and, appropriate cofactors, co-repressors, and co-activators. More specifically, HOXD11 motif discovered here exhibited notable changes in comparison to HOXD11-Homeodomain-DNA binding preference (M F Berger *et al.*, 2008; Jolma *et al.*, 2013). HOXD11 motif identified here looks remarkably similar to HOXA10-PBX4 dimer binding motif, demonstrating that there are specific changes induced by TALE proteins, as captured by this *de novo* motif analysis (Jolma *et al.*, 2015). Furthermore, HOXD11 example indicates that changes in primary HOX motif as defined in this study are likely due to cofactors altering HOX binding site affinity.

However, the investigation of HOX binding also demonstrated that motif change between *in vitro* and *in vivo* studies is not universal to all HOX-TFs. Specifically; the PG13 group appears to have largely unchanged monomer-like motifs enriched in their data, indicating that they might be less

reliant on TALE, or other cofactors. Furthermore, in line with other studies, the HOXA13 and HOXD13 had both their own, and each other's motifs enriched, suggesting they can bind DNA through both motifs (Zhang *et al.*, 2011; Turner *et al.*, 2014). These results clearly suggest that HOX-DNA binding specificities rely on cofactors for indirect but also for direct binding.

5.2.4 Indirect HOX Binding Sites

De novo motif analysis aimed at investigation of indirect binding uncovered three motifs of putative HOX cofactors: AP1, "Unmatched", and CTCF.

"Unmatched" - a Speculative HOXA13 Cofactor

"Unmatched" motif is a CG rich motif absent from public database. It was centrally enriched and abundant in HOXA13 peaks, indicating it could be a potential tethering factor specific for HOXA13. However, due to the inability to match this motif with one protein, it is challenging to speculate on any possible role this factor could play for HOXA13-DNA binding.

AP1 - a General Cofactor

Activating Protein 1 (AP1) is a transcription factor involved in gene activation. It binds the DNA in diverse heterodimers coming from any of the four protein families, cJun, cFos, ATF, or JDP. It is linked to a broad range of developmental and differentiation processes, but so far has not been linked to any HOX-related process (Hess 2004). Here, secondary *de novo* motif search revealed an AP1 motif, indicating AP1 as a possible HOX cofactor. Furthermore, transcriptomic analysis demonstrated a high upregulation of FOS and JUN suggesting a cross-regulatory mechanism. Finally, this motif has also been identified as a cofactor of many other TFs, such as MSX2, RUNX2, and HOXD13^{R298Q} suggesting that AP1 is frequently used as a transcriptional cofactor during developmental processes (Turpaev, 2006; Hein, 2013; Ibrahim, 2014).

CTCF - a HOX Cofactor

Unbiased comparison of HOX binding profiles indicated a subdivision into two groups: Group 1 and Group 2. This subgrouping was in part attributed to abundance of indirect binding, mainly through CTCF. So far, no study linked the role of CTCF and HOX-driven transcriptional programs.

CTCF is an insulator protein and, together with Cohesin, partly controls genome architecture and enhancer-promoter looping (Dixon *et al.*, 2012; Nora *et al.*, 2012; Sexton *et al.*, 2012; Zuin *et al.*,

2013; Sanborn *et al.*, 2015). Furthermore, RAD21 (a Cohesin subunit) was often binding HOX-CTCF co-occupied sites, indicating that triple (HOX-CTCF-RAD21) sites could collectively act in local genome organization, either through insulation, initiation, or stabilization of specific enhancer-promoter contacts (**Figure 5.1**). In this way, CTCF and Cohesin could help secure stable, tissue specific expression through local genome architecture (Faure *et al.*, 2012; Merkenschlager and Odom, 2013; Ing-Simmons *et al.*, 2015). Finally, three possibilities do not exclude one another and are all likely mechanisms of gene regulation on a subset of HOX target genes.

These conclusions are in agreement with Beccari et al. (2016), where HOXA13/HOXD13 loss-of-function mice are unable to switch from early to late *Hoxd* regulation and as a consequence, they have deformed autopod. This deformation is accompanied by an abnormal change in H3K27Ac and H3K27me3 over the HoxD locus, indicating that in this case, HOX13 proteins are essential for the local micro-architecture and chromatin remodeling. Furthermore, Sheth et al. (2016) used the same HOXA13/HOXD13 loss-of-function mutant mouse and performed H3K27me3 and H3K27Ac ChIP-seq. They found changes in H3K27me3 and H3K27Ac profiles genome-wide accompanied with transcriptional changes. Results from Beccari et al. (2016), Sheth et al. (2016), and from this study support the theory, where HOX-TFs play a vital role in the rewiring of local HOX-responsive genome architecture and functional chromatin which in turn instigates transcriptional changes.

However, it is still unclear what the exact mechanism for such micro-architecture establishment is. A recent study in *Drosophila* hinted, there could be a link between developmental TFs and CTCF co-binding and signalling pathways (Van Bortle *et al.*, 2015). Van Bortle et al. (2015) demonstrated that SMAD TFs - effectors of the TGF-β/BMP signalling pathway -bind on a subset of CTCF sites, within TADs, only upon exposure to BMP signaling. Protein-protein interactions fully drive this SMAD-CTCF co-binding, and upon depletion of CTCF, SMADs cannot bind these genomic positions anymore. Interestingly, HOX-TFs are known to co-bind with SMAD proteins, further indicating HOX-CTCF co-binding could be a mechanism of gene regulation for a subset of HOX targets (**Figure 5.1**) (Williams et al. 2005).

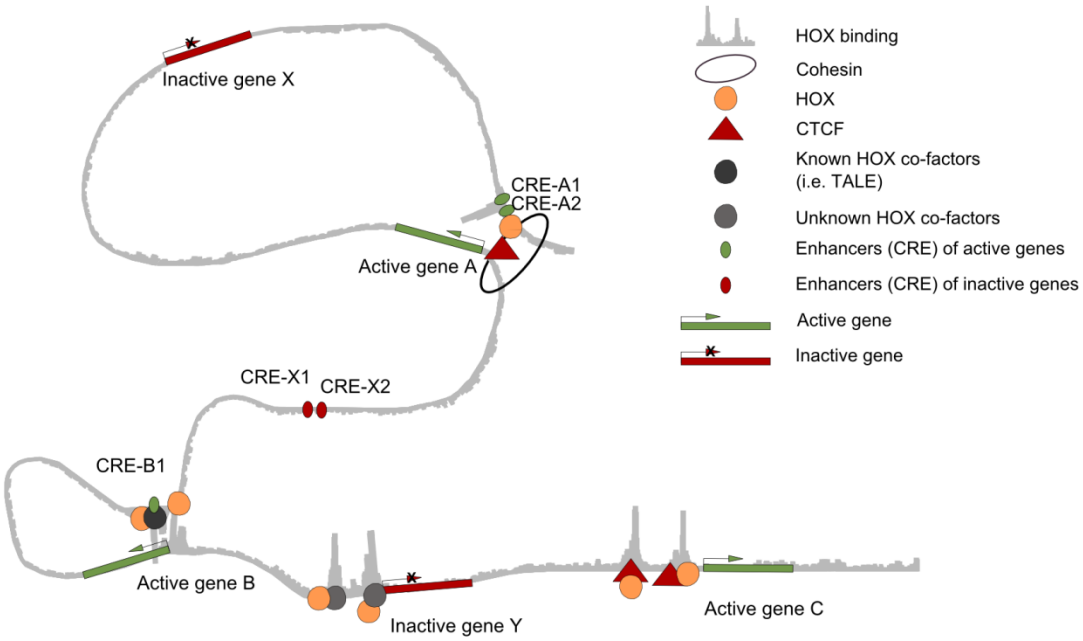


Figure 5.1 Schematic summary of different possibilities of HOX binding events in the genome.
Active gene A: HOX-TFs bind on the CRE-A1 and CRE-A2 and contact CTCF to instigate looping between enhancers and promoter of gene A. At the same time this co-binding insulates Inactive gene X from its own CREs, CRE-X1 and CRE-X2. **Active gene B**: HOX-TFs bind CRE-B1 either as monomer (right) or, co-bind with TALE proteins (left) to activate gene B. **Active gene C**: HOX-TFs either co-bind with CTCF (right) or CTCF tethers HOX proteins to the DNA (left) on proximal CREs to activate gene C. **Inactive gene Y**: HOX-TFs either co-bind with or, are tethered by an unknown cofactor to the proximal CREs of gene Y, thereby

5.3 HOX Induced Restructuring of Functional Chromatin at CTCF sites

Investigation of functional chromatin in HOXA10 and HOXD13 expressing chMM culture demonstrated a gain of H3K27me3, but not H3K4me3 mark around the CTCF binding sites in these cultures.

Narendra et al. (2015) examined the CTCF-induced blockage of H3K4me3 spreading at the HoxA cluster. Indeed, when CTCF sites were deleted, H3K4me3 would spread outside the normally CTCF bound sites. Given the absence of a genome-wide change of H3K4me3 at CTCF sites in HOXA10 and HOXD13 expressing cultures, it is likely that the effect observed by Narendra et al. (2015) is a local feature.

On the other hand, Cuddapah et al. (2009) study focused on the breadth of H3K27me3 mark adjacent to CTCF sites. They noticed that on a genome-wide scale, CTCF demarcates active and repressive domains. Interestingly, in this study, upon HOX overexpression there is a slight gain of H3K27me3 mark at CTCF sites. Here, chromatin remodeling is found at many active promoters belonging to genes involved in metabolic processes, specifically DNA and RNA editing and DNA replication. Therefore, it is not trivial to interpret these findings as regulation of metabolic processes might be linked either to global transcriptional regulation or various homeostatic process.

5.4 HOXA10 Deletion Construct Instability

HOXA10 protein interacts with the CTCF. However, due to the likely protein instability, it is unclear what the exact interaction domain is. PLA assay suggested possiblity that HOXA10 and CTCF bind indirectly. Indeed, in the light of HOX non-canonical (cofactor influenced and tethered) binding strategies, discussed above, it is likely that HOX could be found in a larger complex, perhaps with SMADs and CTCF, as discussed in the **Chapter 5.2.3** (Williams, Williams, Heaton, *et al.*, 2005; Mann, Lelli and Joshi, 2009).

5.5 Outlook

This is the first study of HOX binding *in vivo*. Interestingly, the main conclusions both confirm earlier HOX binding investigations and offer a new perspective on the non-canonical HOX binding sites. HOX-TFs are often bound with cofactors, which alter their sequence affinity. Furthermore, an abundance of HOX binding is attributed to tethering events, which are linked to CTCF and Cohesin. This finding suggests that HOX proteins likely play a role in local genome architecture establishment/management.

It will be very exciting to investigate these sites in more detail, discover mechanistic properties of these binding events, and their functional importance. One possibility of studying this would be by the targeted disruption of CTCF binding motifs underlying co-bound sites and investigation of subsequent (in)ability of HOX-TF to bind to these positions, as well as functional outcome of such mutations. Furthermore, to investigate the local genomic restructuring, it would be ideal to obtain structural 3D chromatin data from wild-type and CTCF mutated motifs and assess

eventual contact rewiring. Lastly, it would be necessary to shed light on the function of these perturbations and thereby, quantify the effect on HOX target gene expression.

The precise mechanistic and functional delineation of HOX-CTCF co-binding could have far reaching implications as it would describe a genuinely novel mechanism for developmental TF target regulation.

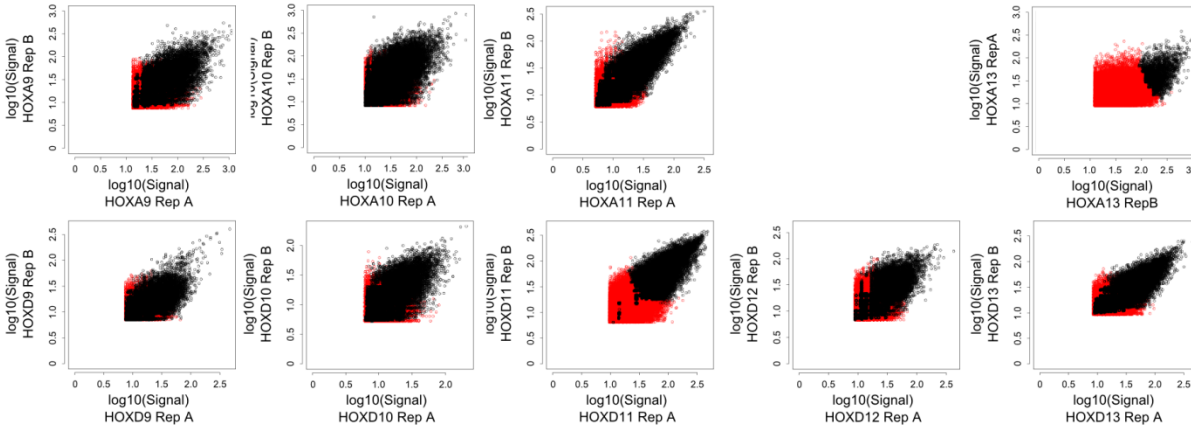
6 Appendix

A

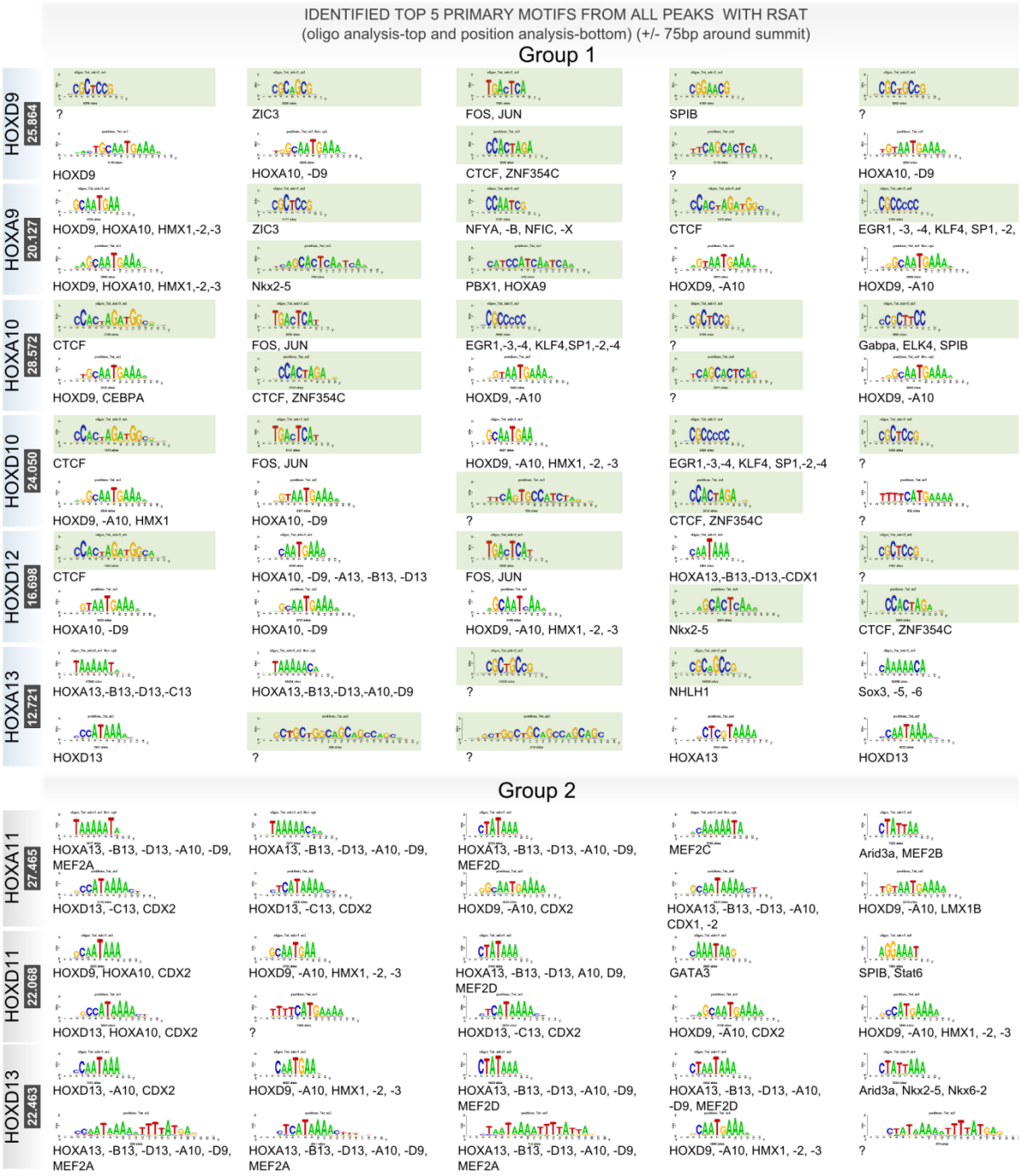
reproducible number of peaks above IDR cutoff 0.01

Sample & IDR cutoff	HOXA9 0.01	HOXA10 0.01	HOXA11 0.01	HOXA13 0.01	HOXD9 0.01	HOXD10 0.01	HOXD11 0.01	HOXD12 0.01	HOXD13 0.01	
RepA	17,101	23,714	23,585	23,471	22,999	18,659	31,190	20,173	23,939	Selfconsitstency
RepB	19,194	27,727	21,640	13,156	22,010	26,783	29,856	28,862	25,448	
RepA vs. RepB	**20,569**	**29,261**	**27,890**	2,569	**26,598**	**24,572**	**22,196**	**17,002**	**22,764**	Consistency between Biological Replicates
Pooled Rep.	24,613	33,262	24,816	**13,236**	25,482	11,361	31,731	25,995	19,933	

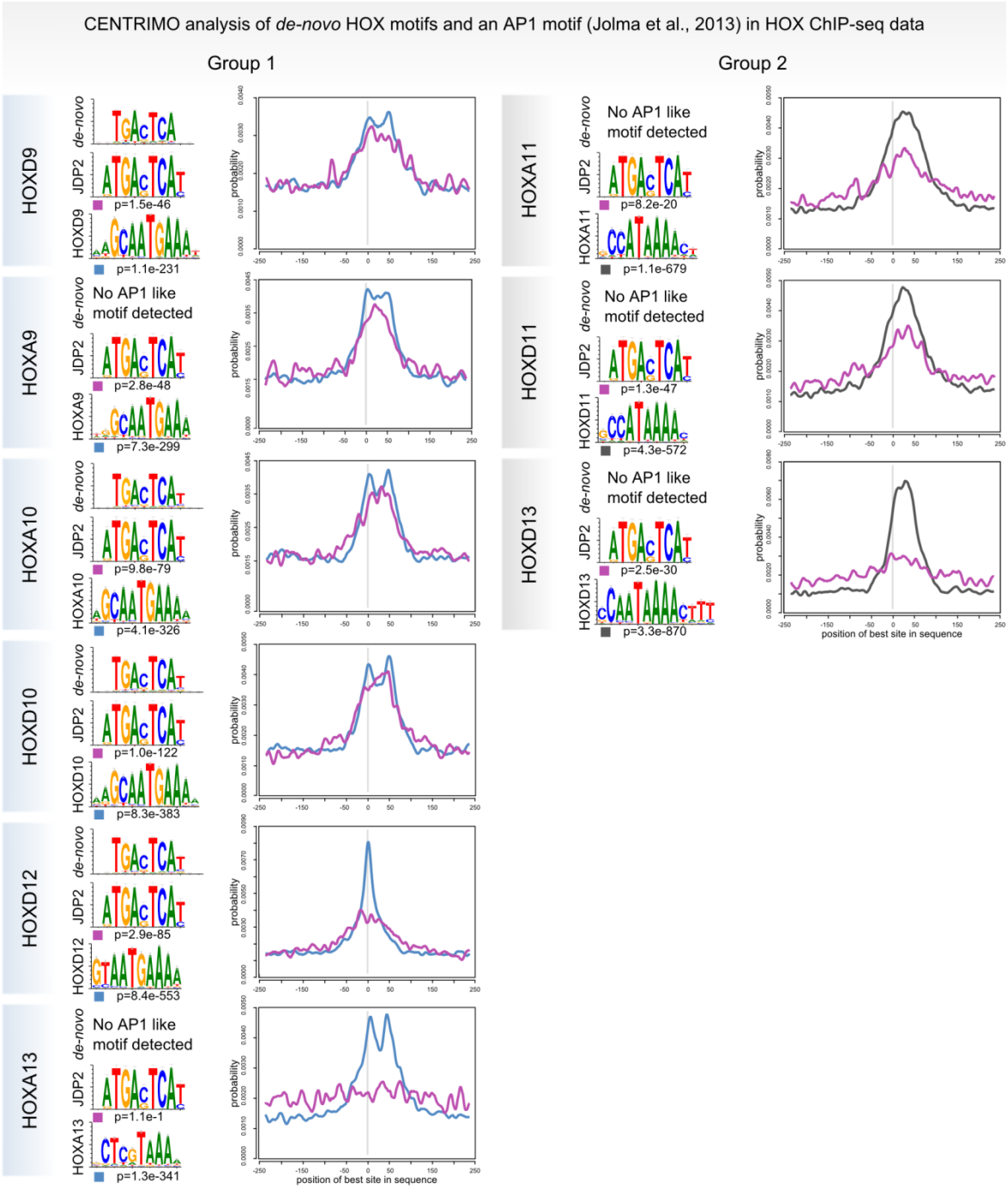
B



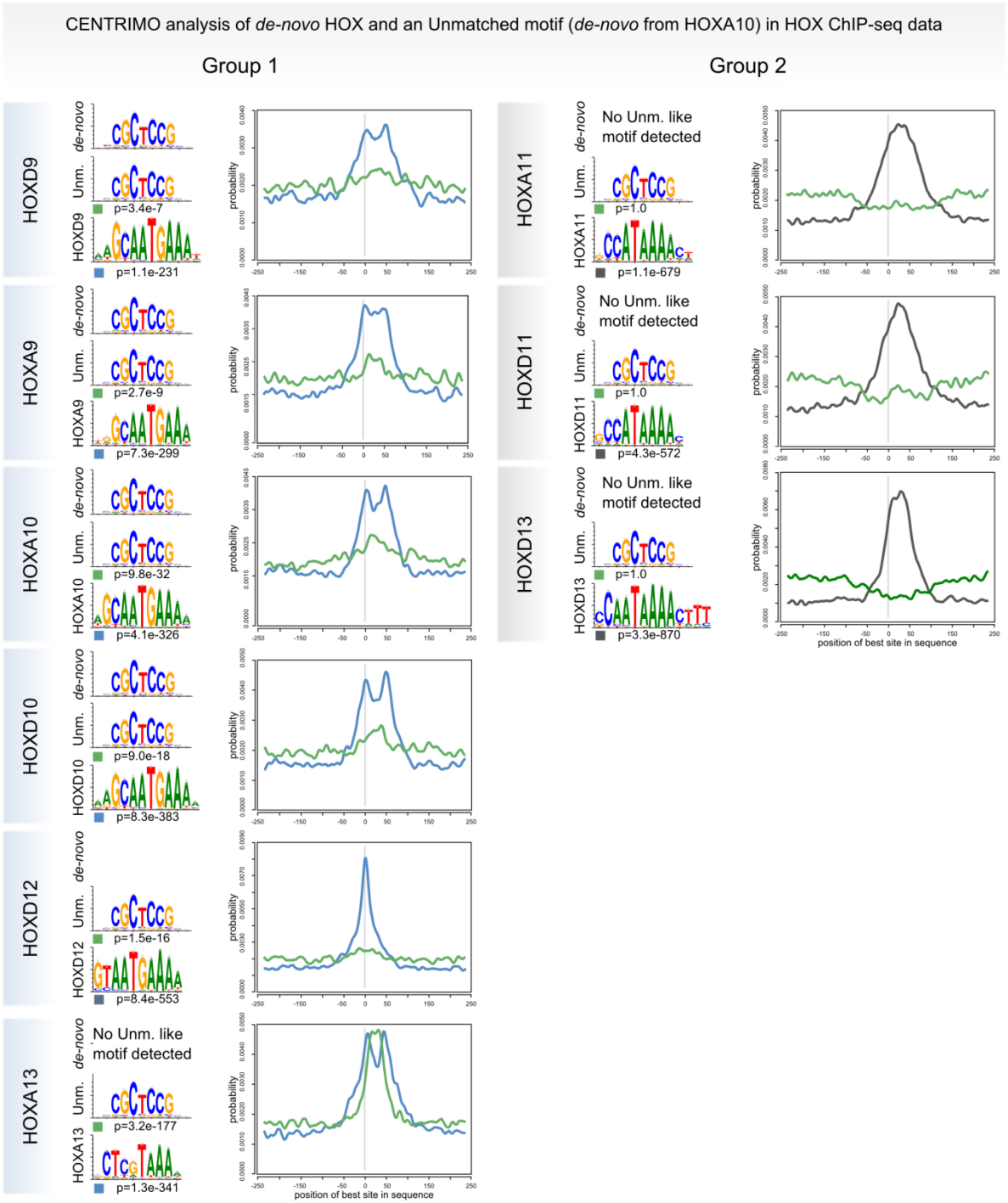
Appendix 1 Summarized Irreproducibility Discovery Rate (IDR) analysis for all the HOX ChIP-seq data.



Appendix 2 All discovered secondary motifs in the HOX ChIP-seq data.



Appendix 3 AP1 Centrimo analysis for all the HOX data.



Appendix 4 "Unmatched" Centrimo analysis for all the HOX data.

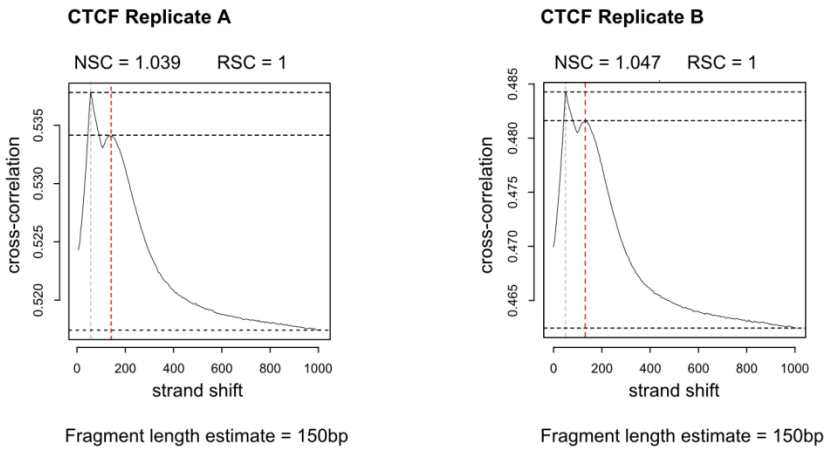
Initial Quality Control and alignment of reads to a reference genome

chCTCF - QC Sheet

Sample (Identifier)	# sequenced reads	# uniquely mapped reads	# non-redundant reads
CTCF - RepA (2014-11-01)	45,812,406	38,068,185	36,703,213
CTCF - RepB (2014-11-01)	37,251,217	30,318,537	29,331,874
Input A - RepA (2014-11-01-1)	40,420,718	33,816,819	32,949,177
Input B - RepB (2014-11-01-1)	41,787,844	34,777,234	33,833,535

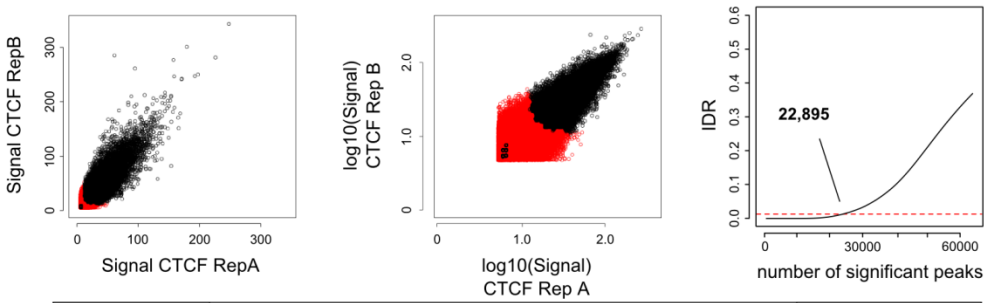
Evaluation of ChIP-efficiency

Cross-correlation Analysis



Validation of Reproducibility

IDR Analysis



Sample	number of peaks above IDR 0.0025	0.005	0.01	0.02	
Replicate 1			14,530	17,486	Selfconsitstency
Replicate 2			17,776	21,139	
Rep1 vs. Rep2	18,125	20,212	**22,895**	26,651	Consitency between Biological Replicates
Pooled Rep.	18,691	20,849	23,443	26,684	

Appendix 5 Quality control and IDR analysis for 3xFLAG-tag CTCF ChIP-seq.

Initial Quality Control and alignment of reads to a reference genome

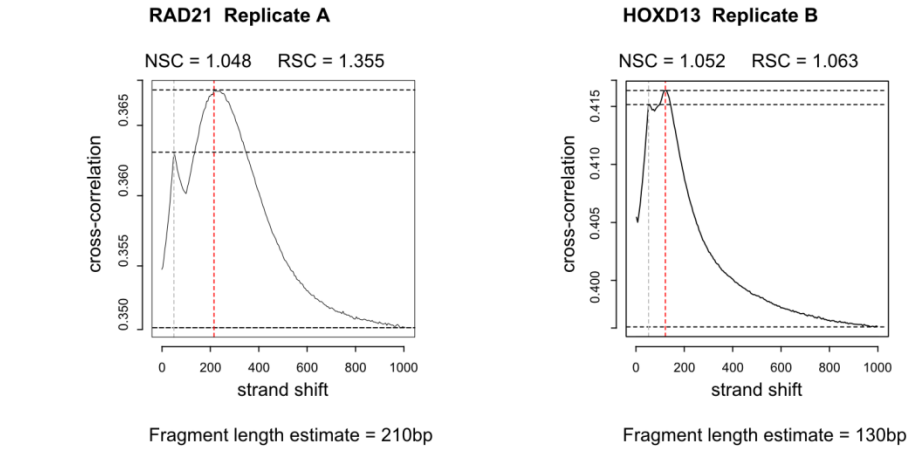
Evaluation of ChIP-efficiency

Validation of Reproducibility

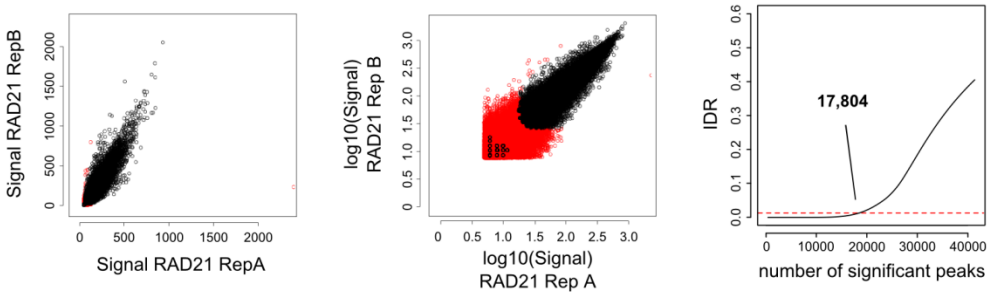
chRAD21 - QC Sheet

Sample (Identifier)	# sequenced reads	# uniquely mapped reads	# non-redundant reads
RAD21 (CTCF infected cultures) - RepA (2014-10-15-11824)	50,928,463	18.319.549	13.898.416
RAD21 (CTCF infected cultures) - RepB (2014-10-15-12213)	16.528.131	13.124.989	10.602.623
Input A (CTCF infected cultures) - input (2014-10-15-3761)	40,420,718	33,816,819	32,949,177
Input B (CTCF infected cultures) - input (2014-10-15-3762)	41,787,844	34,777,234	33,833,535

Cross-correlation Analysis

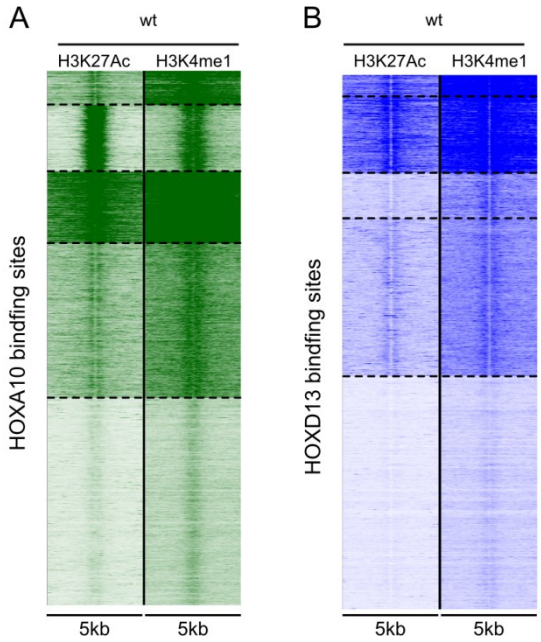


IDR Analysis

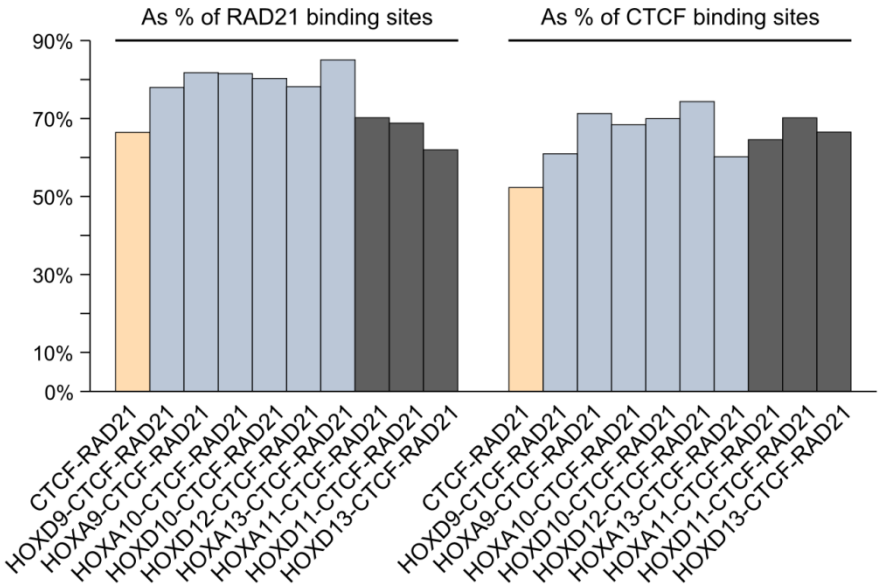


Sample	number of peaks above IDR 0.0025	0.005	0.01	0.02	
Replicate 1			15,675	17,645	Selfconsitstency
Replicate 2			17,859	20,021	
Rep1 vs. Rep2	14,793	16,170	**17,804**	19,791	Consitency between Biological Replicates
Pooled Rep.	16,624	17,941	19,643	21,841	

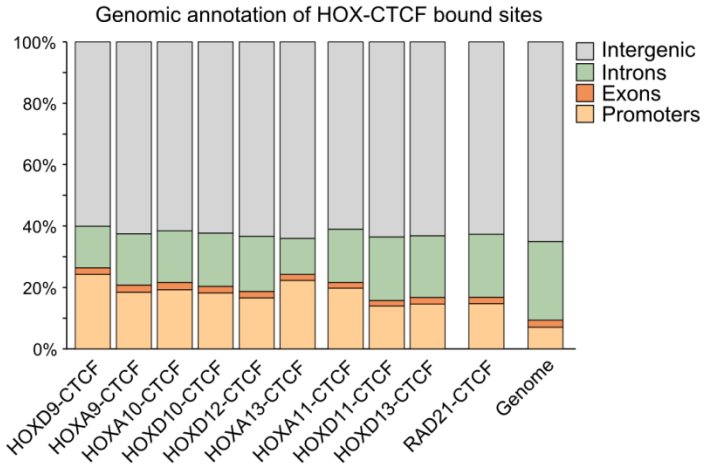
Appendix 6 Quality control and IDR analysis for RAD21 ChIP-seq.



Appendix 8 seqMINER analysis of the H3K27me3 and H3K4me1 binding in the vicinity of HOXA10 and HOXD13 peaks in non-infected chMM.



Appendix 7 Proportion of HOX-CTCF-RAD21 co-bound genomic sites expressed as a percentage of the RAD21 or CTCF binding sites.



Appendix 9 Genomic location of the HOX-CTCF co-bound sites.

7 Literature

Akam, M. (1989) 'Hox and HOM: Homologous gene clusters in insects and vertebrates', *Cell*, pp. 347–349. doi: 10.1016/0092-8674(89)90909-4.

Andrews, N. C. and Faller, D. V. (1991) 'A rapid micrqpreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells', *Nucleic Acids Research*, 19(9), p. 2499. doi: 10.1093/nar/19.9.2499.

Andrey, G. *et al.* (2017) 'Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding', *Genome Research*, 27(2), pp. 223–233. doi: 10.1101/gr.213066.116.

Arnosti, D. N. and Kulkarni, M. M. (2005) 'Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?', *J Cell Biochem*. 2005/02/08, 94(5), pp. 890–898. doi: 10.1002/jcb.20352.

Bailey, T. L. and MacHanick, P. (2012) 'Inferring direct DNA binding from ChIP-seq', *Nucleic acids research*, 40(17), p. e128. doi: 10.1093/nar/gks433.

Banerji, J., Rusconi, S. and Schaffner, W. (1981) 'Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences', *Cell*. 1981/12/01, 27(2 Pt 1), pp. 299–308. doi: 0092-8674(81)90413-X [pii].

Beccari, L. *et al.* (2016) 'A role for HOX13 proteins in the regulatory switch between TADs at the HoxD locus', *Genes and Development*, 30(10), pp. 1172–1186. doi: 10.1101/gad.281055.116.

Berger, M. F. *et al.* (2008) 'Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences', *Cell*. 2008/07/01, 133(7), pp. 1266–1276. doi: S0092-8674(08)00683-1 [pii] 10.1016/j.cell.2008.05.024.

Berger, M. F. *et al.* (2008) 'Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences.', *Cell*, 133(7), pp. 1266–76. doi: 10.1016/j.cell.2008.05.024.

Blankenberg, D. *et al.* (2010) 'Galaxy: A web-based genome analysis tool for experimentalists', *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb1910s89.

Van Bortle, K. *et al.* (2015) 'CTCF-dependent co-localization of canonical Smad signaling factors at architectural protein binding sites in D. melanogaster', *Cell Cycle*, (July), pp. 00–00. doi: 10.1080/15384101.2015.1053670.

Crocker, J. *et al.* (2015) 'Low affinity binding site clusters confer hox specificity and regulatory robustness.', *Cell*, 160(1–2), pp. 191–203. doi: 10.1016/j.cell.2014.11.041.

Cuddapah, S. *et al.* (2009) 'Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains.', *Genome research*, 19(1), pp. 24–32. doi: 10.1101/gr.082800.108.

Davis, A. P. *et al.* (1995) 'Absence of radius and ulna in mice lacking hoxa-11 and hoxd-11', *Nature*. 1995/06/29, 375(6534), pp. 791–795. doi: 10.1038/375791a0.

Davis, A. P. and Capecchi, M. R. (1994) 'Axial homeosis and appendicular skeleton defects in mice with a targeted disruption of hoxd-11', *Development*. 1994/08/01, 120(8), pp. 2187–2198. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7925020.

DeLise, A. M. *et al.* (2000) 'Embryonic Limb Mesenchyme Micromass Culture as an In Vitro Model for Chondrogenesis and Cartilage Maturation', *Methods in Molecular Biology*, 137, pp. 359–375. doi: 10.1385/1-59259-066-7:359.

Deng, W. *et al.* (2012) 'Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor', *Cell*, 149(6), pp. 1233–1244. doi: 10.1016/j.cell.2012.03.051.

Di-Poi, N., Zakany, J. and Duboule, D. (2007) 'Distinct roles and regulations for HoxD genes in metanephric kidney development', *PLoS Genet*. 2007/12/28, 3(12), p. e232. doi: 07-PLGE-RA-0477 [pii] 10.1371/journal.pgen.0030232.

Dixon, J. R. *et al.* (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*. 2012/04/13, 485(7398), pp. 376–380. doi: 10.1038/nature11082 nature11082 [pii].

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dolle, P. *et al.* (1993) 'Disruption of the Hoxd-13 gene induces localized heterochrony leading to mice with neotenic limbs', *Cell*. 1993/11/05, 75(3), pp. 431–441. doi: 0092-8674(93)90378-4 [pii].

Duboule, D. (1994) 'Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony.', *Development (Cambridge, England). Supplement*, 42, pp. 135–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7579514.

Duboule, D. and Morata, G. (1994) 'Colinearity and functional hierarchy among genes of the homeotic complexes', *Trends Genet*. 1994/10/01, 10(10), pp. 358–364. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7985240.

Eden, E. *et al.* (2007) 'Discovering motifs in ranked lists of DNA sequences', *PLoS Computational Biology*, 3(3), pp. 0508–0522. doi: 10.1371/journal.pcbi.0030039.

Eden, E. *et al.* (2009) 'GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists', *BMC Bioinformatics*, 10(1), p. 48. doi: 10.1186/1471-2105-10-48.

Farley, E. K. *et al.* (2015) 'Suboptimization of developmental enhancers', *Science*, 350(6258), pp. 325–328. doi: 10.1126/science.aac6948.

Faure, A. J. *et al.* (2012) 'Cohesin regulates tissue-specific expression by stabilizing highly occupied cis -regulatory modules', (Misteli 2007), pp. 2163–2175. doi: 10.1101/gr.136507.111.Freely.
Featherstone, M. S. *et al.* (1988) 'Hox-5.1 defines a homeobox-containing gene locus on mouse chromosome 2', *Proc Natl Acad Sci U S A*. 1988/07/01, 85(13), pp. 4760–4764. doi: Doi 10.1073/Pnas.85.13.4760.
Fromental-Ramain, C., Warot, X., Messadecq, N., *et al.* (1996) 'Hoxa-13 and Hoxd-13 play a crucial role in the patterning of the limb autopod', *Development*. 1996/10/01, 122(10), pp. 2997–3011. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8898214.
Fromental-Ramain, C., Warot, X., Lakkaraju, S., *et al.* (1996) 'Specific and redundant functions of the paralogous Hoxa-9 and Hoxd-9 genes in forelimb and axial skeleton patterning', *Development*. 1996/02/01, 122(2), pp. 461–472. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8625797.
Garcia-Fernandez, J. (2005) 'The genesis and evolution of homeobox gene clusters', *Nat Rev Genet*. 2005/12/13, 6(12), pp. 881–892. doi: nrg1723 [pii] 10.1038/nrg1723.
Gehring, W. J. *et al.* (1994) 'Homeodomain-DNA recognition', *Cell*, pp. 211–223. doi: 10.1016/0092-8674(94)90292-5.
Gehring, W. J., Kloter, U. and Suga, H. (2009) 'Evolution of the Hox gene complex from an evolutionary ground state.', *Current topics in developmental biology*, 88, pp. 35–61. doi: 10.1016/S0070-2153(09)88002-2.
Genetic Science Learning Center, U. of U. (2016) 'Homeotic Genes and Body Patterns.', *Learn.Genetics*. Available at: http://learn.genetics.utah.edu/content/basics/hoxgenes/.
Goecks, J. *et al.* (2010) 'Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences', *Genome Biology*, 11(8), p. R86. doi: 10.1186/gb-2010-11-8-r86.
Gong, K.-Q. *et al.* (2007) 'A Hox-Eya-Pax Complex Regulates Early Kidney Developmental Gene Expression', *Molecular and Cellular Biology*, 27(21), pp. 7661–7668. doi: 10.1128/MCB.00465-07.
Grant, C. E., Bailey, T. L. and Noble, W. S. (2011) 'FIMO: Scanning for occurrences of a given motif', *Bioinformatics*, 27(7), pp. 1017–1018. doi: 10.1093/bioinformatics/btr064.
Guo, Y. *et al.* (2015) 'CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function', *Cell*. Elsevier Inc., 162(4), pp. 900–910. doi: 10.1016/j.cell.2015.07.038.
Gupta, S. *et al.* (2007) 'Quantifying similarity between motifs', *Genome Biology*, 8(2), p. R24. doi: 10.1186/gb-2007-8-2-r24.
Hein, H. (2013) *Funktionelle Analysen von Transkriptionsfaktoren mit einer Rolle in der chondrogenen und osteogenen Differenzierung mittels ChIP-seq*. Free University Berlin.
Ibrahim, D. M. *et al.* (2013) 'Distinct global shifts in genomic binding profiles of limb malformation-Associated HOXD13 mutations', *Genome Research*, 23(12), pp. 2091–2102. doi: 10.1101/gr.157610.113.
Ibrahim, D. M. (2014) *ChIP-seq Reveals Mutation-Specific Pathomechanisms of HOXD13 Missense Mutations Dissertat ion*.
Ing-Simmons, E. *et al.* (2015) 'Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin', *Genome Research*, 25(4), pp. 504–513. doi: 10.1101/gr.184986.114.
Jeong, Y. *et al.* (2006) 'A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers', *Development*. 2006/01/13, 133(4), pp. 761–772. doi: dev.02239 [pii] 10.1242/dev.02239.
Jolma, A. *et al.* (2013) 'DNA-binding specificities of human transcription factors.', *Cell*. Elsevier Inc., 152(1–2), pp. 327–39. doi: 10.1016/j.cell.2012.12.009.
Jolma, A. *et al.* (2015) 'DNA-dependent formation of transcription factor pairs alters their binding specificity', *Nature*, 527(7578), pp. 384–388. doi: 10.1038/nature15518.
Junion, G. *et al.* (2012) 'A transcription factor collective defines cardiac cell fate and reflects lineage history', *Cell*. 2012/02/07, 148(3), pp. 473–486. doi: S0092-8674(12)00096-7 [pii] 10.1016/j.cell.2012.01.030.
Kessel, M. and Gruss, P. (1990) 'Murine developmental control genes', *Science*. 1990/07/27, 249(4967), pp. 374–379. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1974085.
Kharchenko, P. V, Tolstorukov, M. Y. and Park, P. J. (2008) 'Design and analysis of ChIP-seq experiments for DNA-binding proteins', *Nature biotechnology*, 26(12), pp. 1351–1359. doi: 10.1038/nbt.1508.
Kmita, M. *et al.* (2002) 'Serial deletions and duplications suggest a mechanism for the collinearity of Hoxd genes in limbs.', *Nature*, 420(6912), pp. 145–50. doi: 10.1038/nature01189.
Kmita, M. *et al.* (2005) 'Early developmental arrest of mammalian limbs lacking HoxA/HoxD gene function', *Nature*. 2005/06/24, 435(7045), pp. 1113–1116. doi: nature03648 [pii] 10.1038/nature03648.
Kmita, M. *et al.* (2005) 'Early developmental arrest of mammalian limbs lacking HoxA/HoxD gene function', *Nature*. 2005/06/24, 435(7045), pp. 1113–1116. doi: nature03648 [pii] 10.1038/nature03648.
Kondo, T. *et al.* (1997) 'Of fingers, toes and penises', *Nature*. 1997/11/18, 390(6655), p. 29. doi: 10.1038/36234.
Kulkarni, M. M. and Arnosti, D. N. (2003) 'Information display by transcriptional enhancers', *Development*. 2003/12/09, 130(26), pp. 6569–6575. doi: 10.1242/dev.00890 130/26/6569 [pii].
Kuss, P. *et al.* (2009) 'Mutant Hoxd13 induces extra digits in a mouse model of synpolydactyly directly and by decreasing retinoic acid synthesis', 119(1). doi: 10.1172/JCI36851.146.
Kvon, E. Z. *et al.* (2016) 'Progressive Loss of Function in a Limb Enhancer during Snake Evolution', *Cell*, 167(3), p.

633–642.e11. doi: 10.1016/j.cell.2016.09.028.
Landt, S. G. *et al.* (2012) 'ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.', *Genome research*, 22(9), pp. 1813–31. doi: 10.1101/gr.136184.111.
LaRonde-LeBlanc, N. a and Wolberger, C. (2003) 'Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior.', *Genes & development*, 17(16), pp. 2060–72. doi: 10.1101/gad.1103303.
Lee, T. I., Johnstone, S. E. and Young, R. A. (2006) 'Chromatin immunoprecipitation and microarray-based analysis of protein location', *Nat Protoc*. 2007/04/05, 1(2), pp. 729–748. doi: nprot.2006.98 [pii] 10.1038/nprot.2006.98.
Lee, T. I. and Young, R. A. (2000) 'Transcription of Eukaryotic Protein-Coding Genes', *Annual Review of Genetics*, 34(1), pp. 77–137. doi: 10.1146/annurev.genet.34.1.77.
Lettice, L. A. *et al.* (2003) 'A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly', *Hum Mol Genet*, 12(14), pp. 1725–1735. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12837695.
Lewis, E. B. (1978) 'A gene complex controlling segmentation in Drosophila', *Nature*. 1978/12/07, 276(5688), pp. 565–570. Available at: http://www.ncbi.nlm.nih.gov/pubmed/103000.
Li, H. *et al.* (2009) 'The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup', *Bioinformatics (Oxford, England)*, 25(16), pp. 1–2. doi: 10.1093/bioinformatics/btp352.
Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.
Li, Q. *et al.* (2011) 'Measuring reproducibility of high-throughput experiments', *Annals of Applied Statistics*, 5(3), pp. 1752–1779. doi: 10.1214/11-AOAS466.
Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.
Löytynoja, A. and Goldman, N. (2010) 'webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser', *BMC Bioinformatics*, 11(1), p. 579. doi: 10.1186/1471-2105-11-579.
Mann, R. S., Lelli, K. M. and Joshi, R. (2009) 'Chapter 3 Hox Specificity. Unique Roles for Cofactors and Collaborators', *Current Topics in Developmental Biology*, pp. 63–101. doi: 10.1016/S0070-2153(09)88003-4.
Mathelier, A. *et al.* (2014) 'JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt997.
McLean, C. Y. *et al.* (2010) 'GREAT improves functional interpretation of cis-regulatory regions', *Nature Biotechnology*, 28(5), pp. 495–501. doi: 10.1038/nbt.1630.
Merabet, S. and Lohmann, I. (2015) 'Toward a new twist in Hox and TALE DNA-binding specificity.', *Developmental cell*. Elsevier Inc., 32(3), pp. 259–61. doi: 10.1016/j.devcel.2015.01.030.
Merabet, S. and Mann, R. S. (2016) 'To Be Specific or Not: The Critical Relationship Between Hox And TALE Proteins', *Trends in Genetics*, pp. 334–347. doi: 10.1016/j.tig.2016.03.004.
Merika, M. *et al.* (1998) 'Recruitment of CBP/p300 by the IFN\textbeta\ enhanceosome is required for synergistic activation of transcription.', *Molecular Cell*, 1(2), pp. 277–287. doi: 10.1016/S1097-2765(00)80028-3.
Merika, M. and Thanos, D. (2001) 'Enhanceosomes', *Curr Opin Genet Dev*. 2001/03/16, 11(2), pp. 205–208. doi: S0959-437X(00)00180-5 [pii].
Merkenschlager, M. and Odom, D. T. (2013) 'CTCF and cohesin: Linking gene regulatory elements with their targets', *Cell*, pp. 1285–1297. doi: 10.1016/j.cell.2013.02.029.
Moreau, P. *et al.* (1981) 'The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants', *Nucleic Acids Res*. 1981/11/25, 9(22), pp. 6047–6068. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6273820.
Morgan, B. A. and Fekete, D. M. (1996) 'Manipulating Gene Expression with Replication-Competent Retroviruses', *Methods in Cell Biology*, 51, pp. 185–218. doi: doi.org/10.1016/S0091-679X(08)60629-9.
Mukherjee, K. and Bürglin, T. R. (2007) 'Comprehensive analysis of animal TALE homeobox genes: New conserved motifs and cases of accelerated evolution', *Journal of Molecular Evolution*, 65(2), pp. 137–153. doi: 10.1007/s00239-006-0023-0.
Narendra, V. *et al.* (2015) 'CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation', *Science*, 347(6225), pp. 1017–1021. doi: 10.1126/science.1262088.
Narendra, V. *et al.* (2015) 'Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation.', *Science (New York, N.Y.)*, 347(6225), pp. 1017–21. doi: 10.1126/science.1262088.
Nelson, C. E. *et al.* (1996) 'Analysis of Hox gene expression in the chick limb bud.', *Development (Cambridge, England)*, 122(5), pp. 1449–66. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8625833.
Nobrega, M. A. and Pennacchio, L. A. (2004) 'Comparative genomic analysis as a tool for biological discovery', *The Journal of Physiology*, 554(1), pp. 31–39. doi: 10.1113/jphysiol.2003.050948.
Nora, E. P. *et al.* (2012) 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature*, 485(7398), pp. 381–385. doi: 10.1038/nature11049.
Nora, E. P. *et al.* (2016) 'Targeted degradation of CTCF decouples local insulation of chromosome domains from higher-order genomic compartmentalization', *bioRxiv*, p. 95802. doi: https://doi.org/10.1101/095802.

Noyes, M. B. *et al.* (2008) 'Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites', *Cell*. 2008/07/01, 133(7), pp. 1277–1289. doi: S0092-8674(08)00682-X [pii] 10.1016/j.cell.2008.05.023.
Ong, C.-T. and Corces, V. G. (2014) 'CTCF: an architectural protein bridging genome topology and function.', *Nature reviews. Genetics*. Nature Publishing Group, 15(4), pp. 234–46. doi: 10.1038/nrg3663.
Orgeur, M. (2016) *Transcriptional regulatory network underlying connective tissue differentiation during limb development*. Universit´e Pierre et Marie Curie - Paris VI. Available at: https://tel.archives-ouvertes.fr/tel-01474874.
Pascual-Anaya, J. *et al.* (2013) 'Evolution of Hox gene clusters in deuterostomes', *BMC Developmental Biology*, 13(1), p. 26. doi: 10.1186/1471-213X-13-26.
Phillips-Cremins, J. E. and Corces, V. G. (2013) 'Chromatin Insulators: Linking Genome Organization to Cellular Function', *Molecular Cell*, pp. 461–474. doi: 10.1016/j.molcel.2013.04.018.
Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: A flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.
Raines, A. M. *et al.* (2015) 'Key pathways regulated by HoxA9,10,11/HoxD9,10,11 during limb development.', *BMC developmental biology*. BMC Developmental Biology, 15, p. 28. doi: 10.1186/s12861-015-0078-5.
Rao, S. S. *et al.* (2014) 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*, 159(7), pp. 1665–1680. doi: 10.1016/j.cell.2014.11.021.
Roeder, R. G. (1996) 'The role of general initiation factors in transcription by RNA polymerase II.', *Trends in biochemical sciences*, 21(September), pp. 327–335. doi: http://dx.doi.org/10.1016/S0968-0004(96)10050-5.
Sagai, T. *et al.* (2004) 'Phylogenetic conservation of a limb-specific, cis-acting regulator of Sonic hedgehog (Shh)', *Mamm Genome*. 2004/01/17, 15(1), pp. 23–34. doi: 10.1007/s00335-033-2317-5.
Sambrook, J. and Russell, D. W. (2001) 'Molecular Cloning - Sambrook & Russel - Vol. 1, 2, 3', *Cold Spring Harbor Laboratory Press*, 3th Editio. doi: 10.1002/humu.1186.abs.
Sanborn, A. L. *et al.* (2015) 'Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes', *Proc Natl Acad Sci U S A*, 112(47), pp. E6456--65. doi: 10.1073/pnas.1518552112.
Schwarzer, W. *et al.* (2016) 'Two independent modes of chromosome organization are revealed by cohesin removal', *bioRxiv*, p. 94185. doi: 10.1101/094185.
Seemann, P. (2006) *Zur Bedeutung des Wachstumsfaktors GDF5*. Frei Universität Berlin. Available at: http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000002110.
Sexton, T. *et al.* (2012) 'Three-dimensional folding and functional organization principles of the Drosophila genome', *Cell*, pp. 458–472. doi: 10.1016/j.cell.2012.01.010.
Shen, W. -f. *et al.* (2001) 'The HOX Homeodomain Proteins Block CBP Histone Acetyltransferase Activity', *Molecular and Cellular Biology*, 21(21), pp. 7509–7522. doi: 10.1128/MCB.21.21.7509-7522.2001.
Sheth, R. *et al.* (2016) 'Distal Limb Patterning Requires Modulation of cis-Regulatory Activities by HOX13', *Cell Reports*, 17(11), pp. 2913–2926. doi: 10.1016/j.celrep.2016.11.039.
Slattery *et al.* (2011) 'Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins.', *Cell*. Elsevier Inc., 147(6), pp. 1270–82. doi: 10.1016/j.cell.2011.10.053.
Slattery, M. *et al.* (2011) 'Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins', *Cell*. 2011/12/14, 147(6), pp. 1270–1282. doi: S0092-8674(11)01370-5 [pii] 10.1016/j.cell.2011.10.053.
Slattery, M. *et al.* (2011) 'Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins.', *Cell*. 2011/12/14. Elsevier Inc., 147(6), pp. 1270–1282. doi: 10.1016/j.cell.2011.10.053.
Slattery, M. *et al.* (2014) 'Absence of a simple code: How transcription factors read the genome', *Trends in Biochemical Sciences*, pp. 381–399. doi: 10.1016/j.tibs.2014.07.002.
Small, K. M. and Potter, S. S. (1993) 'Homeotic transformations and limb defects in Hox A11 mutant mice', *Genes and Development*, 7(12 A), pp. 2318–2328. doi: 10.1101/gad.7.12a.2318.
Spitz, F. and Furlong, E. E. (2012) 'Transcription factors: from enhancer binding to developmental control', *Nat Rev Genet*. 2012/08/08, 13(9), pp. 613–626. doi: 10.1038/nrg3207 nrg3207 [pii].
Taneda, S. *et al.* (2004) 'Separation and characterization of alkyltrimethylbenzene derivatives in diesel exhaust particles (DEP)', *Environ Sci*. 2005/03/08, 11(3), pp. 171–178. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15750584.
Thanos, D. and Maniatis, T. (1995) 'Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome', *Cell*. 1995/12/29, 83(7), pp. 1091–1100. doi: 0092-8674(95)90136-1 [pii].
Thomas-Chollier, M. *et al.* (2012) 'A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs.', *Nature protocols*, 7(8), pp. 1551–68. doi: 10.1038/nprot.2012.088.
Turner, M. *et al.* (2014) 'Chemical shift assignments of mouse HOXD13 DNA binding domain bound to duplex DNA', *Biomolecular NMR Assignments*, 9(2), pp. 267–270. doi: 10.1007/s12104-014-9589-4.
Turpaev, K. T. (2006) 'Role of transcription factor AP-1 in integration of cell signaling systems', *Molecular Biology*, 40(6), pp. 851–866. doi: 10.1134/S0026893306060033.
UCSF (2015) *Orthoretriever*. Available at: https://lighthouse.ucsf.edu/orthoretriever/%0A.
Visel, A. *et al.* (2007) 'VISTA Enhancer Browser - A database of tissue-specific human enhancers', *Nucleic Acids*

Research, 35(SUPPL. 1). doi: 10.1093/nar/gkl822.
Wagner, G. P., Amemiya, C. and Ruddle, F. (2003) 'Hox cluster duplications and the opportunity for evolutionary novelties', *Proceedings of the National Academy of Sciences*, 100(25), pp. 14603–14606. doi: 10.1073/pnas.2536656100.
Wellik, D. M. and Capecchi, M. R. (2003) 'Hox10 and Hox11 genes are required to globally pattern the mammalian skeleton', *Science*. 2003/07/19, 301(5631), pp. 363–367. doi: 10.1126/science.1085672 301/5631/363 [pii].
Williams, T. M., Williams, M. E., Kuick, R., *et al.* (2005) 'Candidate downstream regulated genes of HOX group 13 transcription factors with and without monomeric DNA binding capability', *Developmental Biology*, 279(2), pp. 462–480. doi: 10.1016/j.ydbio.2004.12.015.
Williams, T. M., Williams, M. E., Heaton, J. H., *et al.* (2005) 'Group 13 HOX proteins interact with the MH2 domain of R-Smads and modulate Smad transcriptional activation functions independent of HOX DNA-binding capability', *Nucleic Acids Research*, 33(14), pp. 4475–4484. doi: 10.1093/nar/gki761.
de Wit, E. *et al.* (2015) 'CTCF Binding Polarity Determines Chromatin Looping', *Molecular Cell*, 60(4), pp. 676–684. doi: 10.1016/j.molcel.2015.09.023.
Wolpert, L., Tickle, C. and Martinez, A. M. (2015) *Principles of development*, *Book*. Available at: http://discovery.ucl.ac.uk/21856/.
Woltering, J. M. and Duboule, D. (2010) 'The origin of digits: expression patterns versus regulatory mechanisms', *Dev Cell*. 2010/04/24, 18(4), pp. 526–532. doi: 10.1016/j.devcel.2010.04.002 S1534-5807(10)00153-X [pii].
Ye, T. *et al.* (2011) 'seqMINER: an integrated ChIP-seq data interpretation platform.', *Nucleic acids research*, 39(6), p. e35. doi: 10.1093/nar/gkq1287.
Zakany, J. and Duboule, D. (2007) 'The role of Hox genes during vertebrate limb development', *Curr Opin Genet Dev*. 2007/07/24, 17(4), pp. 359–366. doi: S0959-437X(07)00116-5 [pii] 10.1016/j.gde.2007.05.011.
Zakany, J. and Duboule, D. (2007) 'The role of Hox genes during vertebrate limb development', *Curr Opin Genet Dev*. 2007/07/24, 17(4), pp. 359–366. doi: S0959-437X(07)00116-5 [pii] 10.1016/j.gde.2007.05.011.
Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS).', *Genome Biology*, 9(9), p. R137. doi: 10.1186/gb-2008-9-9-r137.
Zhang, Y. *et al.* (2011) 'Structural basis for sequence specific DNA binding and protein dimerization of HOXA13', *PLoS ONE*, 6(8). doi: 10.1371/journal.pone.0023069.
Zhao, Y. and Potter, S. S. (2002) 'Functional Comparison of the Hoxa 4, Hoxa 10, and Hoxa 11 Homeoboxes', *Developmental Biology*, 244(1), pp. 21–36. doi: 10.1006/dbio.2002.0595.
Zuin, J. *et al.* (2013) 'Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells', *Proceedings of the National Academy of Sciences*, 111(3), pp. 996–1001. doi: 10.1073/pnas.1317788111.

8 Acknowledgements

First, I would like to thank my supervisor Prof.Stefan Mundlos for allowing me to work in his group, for supporting me for so long, and for allowing me to freely pursue my scientific curiosity. I am forever indebted for the opportunity to work in, by all accounts, a fantastic group. Next, I would like to thank my mentor Dr. Jochen Hecht for his guidance, support, and for the freedom to follow my own ideas.

Of course, I would be nowhere if not for the support of the best colleagues one could ever ask for. The entire AG Mundlos is not just a workplace but a place of creativity, discussions, humor, and friendship. You make work feel like a vacation and hard work like a hobby. This group had had an immense impact on me personally and professionally, and I would like to thank all of you for that. Foremost, Daniel — my scientific big brother — I would like to thank you for always having a good piece of advice, your help with my project in so many different ways, and for always being honest with me, even when it is not what I wanted to hear. This has helped me grow and better myself. Guillaume, you are a textbook picture of what a colleague should be. Thank you for all your guidance, discussions, coffees, wine glasses, and fondue. Katerina, my bureaucracy translator, party follower, and intense discussion co-participant I truly appreciate all your help and good times we had, both in the lab and outside. Fany and Christina you made our "Jerko" office feel like a small piece of heaven with your love, care, support, and genuine friendship. I am so happy to have had friends around when facing all the hurdles of my thesis. Iza, Georgie, and Aru…what can I say, whatever cosmic event put us at the same time at the same place; it had a great plan. You have been amazing friends, and colleagues at some point. Always on point with your advice and ready with your remedy for hard times, and you know our remedy is Aperol and wine! Sala, how many times can I thank you for driving me home from the MPI? Well, this is now an official thank you, not only for that but for being the awesome self you are. Mike, Martin, Giulia, Dario, Lila, Alex, Bjørt thank you for all the good times we had together, and you know there were more than a few…and Bjørt get a better table for the next flat! Asita thank you for listening to me so often, for your help and support in the lab, I really appreciate it.

Finally, I would like to thank my family for always supporting me and believing that I can do anything I set my mind to. Mom, dad, Nera, and Stipan you give me strength, resilience, and

determination to boldly go further. I will never be able to put in words how much it means to me!

9 Declaration of Independent Work

I hereby declare that I have independently conceived, prepared, carried out, analyzed, and written here presented results and ideas, other than those properly referenced.

No collaboration with the commercial doctoral degree supervisors took place during this work.

Currently I do not hold a PhD degree, nor have I applied for one at any other university.

The principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

This work or any part of it, has not been submitted to, approved, or rejected by any other any academic institution in Germany or abroad.

I declare that I am aware of the doctoral regulations of the Lebenswissenschaftliche Fakultät at the Humboldt-Universität zu Berlin.

_____

Berlin,

10 Publications & Presentations

ORAL & POSTER PRESENTATIONS

2016 Jun	**Poster presentation**: "Cell Symposium: Transcription and Development and Disease" Chicago, IL, USA
2016 May	**Oral presentation**: "Institute seminar series at the Max-Planck Institute for Molecular Genetics" Berlin, Germany
2015 Oct	**Poster presentation**: "EMBO meeting-Transcriptional control in Development and Evolution" Paris, France
2014 Sep	**Poster presentation**: "AGD" Potsdam, Germany
2014 Sep	**Poster presentation**: "Epigenetics & Chromatin" Cold Spring Harbor, NY, USA

ACADEMIC PUBLICATIONS

Genome-wide binding of posterior HOXA/D transcription factors reveals subgrouping and association with CTCF
Jerković I., Ibrahim D.M. , Andrey G., Haas S. , Hansen P., Janetzki C. , Navarrete I.G. , Robinson P.N. , Hecht J., Mundlos S.; PlosGenetics (2017) doi:10.1371/journal.pgen.1006567

Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding
Andrey G.♯, Schöpflin R.♯, **Jerković I**., Heinrich V., Ibrahim DM., Paliou C., Hochradel M., Timmermann B., Vingron M. and Mundlos S.; Genome Research (2016) doi: 10.1101/gr.213066.116

Formation of novel chromatin domains determines pathogenicity of genomic duplications
Franke M. ♯, Ibrahim D.M. ♯, Andrey G., Schwarzer W., Heinrich V., Schöpflin R., Kraft K., Kempfer R., **Jerković I.,** Chan W.L., Spielmann M., Timmermann B., Wittler L., I Kurth., Cambiaso P., Zuffardi O., Houg G., Lambie L., Brancati F., Pombo A., Vingron M., Spitz F., Mundlos S.; Nature (2016) doi:10.1038/nature19800